Grazer Philosophische Studien, Volume 90 – 2014

Grazer Philosophische Studien

Founded by

Rudolf Haller

Editors

Johannes L. Brandl (*Universität Salzburg*)
Marian David (*Universität Graz*)
Maria E. Reicher (*Universität Aachen*)
Leopold Stubenberg (*University of Notre Dame*)

Managing Editor

Martina Fürst (*Universität Graz*)

Editorial Board

Peter Baumann – Monika Betzler – Victor Caston – Annalisa Coliva – Thomas Crisp – Dagfinn Føllesdal – Volker Gadenne – Christopher Gauker – Hanjo Glock – Robert M. Harnish† – Reinhard Kamitz – Thomas Kelly – Andreas Kemmerling – Jaegwon Kim – Peter Koller – Wolfgang Künne – Karel Lambert – Keith Lehrer – Hannes Leitgeb – Joseph Levine – Georg Meggle – Thomas Mormann – Edgar Morscher – Herlinde Pauer-Studer – Christian Piller – Marga Reimer – Edmund Runggaldier – Heiner Rutte – Werner Sauer – Alfred Schramm – Gerhard Schurz – Geo Siegwart – Peter Simons – Barry Smith – Thomas Spitzley – Matthias Steup – Mark Textor – Thomas Uebel – Ted Warfield – Charlotte Werndl – Nicholas White

VOLUME 90

The titles published in this series are listed at *brill.com/gps*

Grazer Philosophische Studien
Volume 90 – 2014

International Journal for Analytic Philosophy

BRILL
RODOPI

LEIDEN | BOSTON

Die Herausgabe der GPS erfolgt mit Unterstützung der Geisteswissenschaftlichen Fakultät der Universität Graz.

Library of Congress Control Number: 2015938342

ISSN 0165-9227
ISBN 978-90-04-29873-6 (paperback)
ISBN 978-90-04-29876-7 (e-book)

This book is printed on acid-free paper.

INHALTSVERZEICHNIS TABLE OF CONTENTS

Themenschwerpunkt / Special Topic
THE SECOND-PERSON STANDPOINT IN LAW AND MORALITY

Gastherausgeber / Guest Editors
Christoph HANISCH & Herlinde PAUER-STUDER

Christoph HANISCH & Herlinde PAUER-STUDER: Editorial 1

Stephen DARWALL: Why Fichte's Second-Personal Foundations Can Provide a More Adequate Account of the Relation of Right than Kant's 5

Fabienne PETER: Second-Personal Reason-Giving 21

Hans Bernhard SCHMID: Missing the "We" for all those "You's". Debunking Milgram's *Obedience to Authority* 35

Peter SCHABER: Demanding Something 63

Alexandra COUTO: Reactive Attitudes, Disdain and the Second-Person Standpoint 79

Christoph HANISCH: Self-Constitution and Other-Constitution: The Non-Optionality of the Second-Person Standpoint 105

Jens TIMMERMANN: Kant and the Second-Person Standpoint 131

Herlinde PAUER-STUDER: Contractualism and the Second-Person Moral Standpoint 149

Abhandlungen *Articles*

Michele PAOLINI PAOLETTI: Falsemakers: Something Negative about Facts 169

Francesco BERTO & Graham PRIEST: Modal Meinongianism and Characterization. Reply to Kroon 183

Shane RYAN: Standard Gettier Cases: A Problem for Greco? 201

Simon DIERIG: The Discrimination Argument and the Standard Strategy 213

Federico CASTELLANO: Intellectualism Against Empiricism 231

Uku TOOMING: Pleasures of the Communicative Conception . . . 253

Uwe PETERS: Teleosemantics, Swampman, and Strong Representationalism 273

Paula SWEENEY: Contextualism and the Principle of Tolerance 289

Essay-Wettbewerb *Essay Competition*

Tammo LOSSAU: Was heißt „sich vorstellen, eine andere Person zu sein“? 307

Eva BACKHAUS: Essay zur Frage: *Kann ich mir vorstellen, eine andere Person zu sein?* 317

Viktoria KNOLL: Kann ich mir vorstellen, eine andere Person zu sein? 327

Besprechung *Review Article*

Guido MELCHIOR: Is Epistemological Disjunctivism the Holy Grail? 335

Buchnotizen *Critical Notes*

Stefania CENTRONE (Hg.), *Versuche über Husserl*. Hamburg: Felix Meiner Verlag, 2013 (Edgar MORSCHER) 347

Edgar MORSCHER: *Normenlogik. Grundlagen – Systeme – Anwendungen*. Paderborn: mentis Verlag, 2012 (Hans-Peter LEEB) 354

Lisa HERZOG, *Inventing the Market: Smith, Hegel, and Political Theory*. Oxford: Oxford University Press. 2013. – Lisa HERZOG and Axel HONNETH (eds.), *Der Wert des Marktes: Ein ökonomisch-philosophischer Diskurs vom 18. Jahrhundert bis zur Gegenwart*. Berlin: Suhrkamp. 2014. (Norbert PAULO) 356

Grazer Philosophische Studien
90 (2014), 1–3.

"THE SECOND-PERSON STANDPOINT IN LAW AND MORALITY"

EDITORIAL

Christoph HANISCH & Herlinde PAUER-STUDER

The papers of this special issue are the outcome of a two-day conference entitled "The Second-Person Standpoint in Law and Morality" that took place at the University of Vienna in March 2013 and was organized by the ERC Advanced Research Grant "Distortions of Normativity".

The aim of the conference was to explore and discuss Stephen Darwall's innovative and influential second-personal account of foundational moral concepts such as "obligation", "responsibility", and "rights", as developed in his book *The Second-Person Standpoint: Morality, Respect, and Accountability* (Harvard University Press 2006) and further elaborated in *Morality, Authority and Law: Essays in Second-Personal Ethics I* and *Honor, History, and Relationships: Essays in Second-Personal Ethics II* (both Oxford University Press 2013).

With the second-person standpoint Darwall refers to the unique *conceptual normative space* that practical deliberators and agents occupy when they *address claims and demands* to one another (and to themselves). The very first sentence of Darwall's examination of the second-personal conceptual paradigm summarizes the gist of the argument succinctly when he claims that "*the second-person standpoint* [is] the perspective that you and I take up when we make and acknowledge claims on one another's conduct and will" (Darwall 2006, 3). *The Second-Person Standpoint* reminds us that this perspective has been ignored for much too long and that it better take centre stage in any philosophical analysis of moral phenomena, in order to yield a satisfying account of morality as a social institution. The negative part of Darwall's strategy is to show that neither a purely first-personal approach (represented by Kant and contemporary Kantians) nor a third-personal state-of-affairs-perspective (represented by most varieties of contemporary consequentialism) are capable of accounting for the categorical bindingness characteristic of moral obligation. The latter feat can only be

accomplished, and this is the positive part of Darwall's argument, when those second-personal normative "felicity conditions" and conceptual presuppositions are acknowledged and spelled out that are already presupposed in *every* instance of issuing (putatively valid) claims and demands. It is especially second-personal competence and second-personal authority that are the bedrock of these normative conceptual presuppositions, without which engaging in any meaningful address would be impossible. Kantians and utilitarians alike have neglected this critical dimension of the normative landscape.

In addition to working out an original conception of moral obligation, the first eight chapters of *The Second-Person Standpoint* articulate this fundamental insight with respect to a variety of traditional projects in ethical theory such as developing accounts of moral responsibility, rights, dignity, and autonomy. In this context, special emphasis is to be awarded, on the one hand, to Darwall's refreshing second-personal interpretation of Strawson's influential account of reactive attitudes and moral responsibility and, on the other, to his historically well-informed reconstruction of Samuel Pufendorf's often neglected version of an enlightened theistic voluntarism concerning moral authority.

Darwall dedicates the second part of *The Second-Person Standpoint* to the urgent question: how should one respond to the sceptical challenge that expresses utter indifference to the second-person standpoint, including all its multifarious normative presuppositions and implications? What commits us to all this? It is at this point that Darwall, firstly, refines his criticisms of the Kantian, first-personal, paradigm of normativity and emphasizes that only if one already incorporates the second-personal conceptual apparatus into a Kantian analysis of moral obligation is the latter going to yield a convincing account. Secondly, and this certainly is one of the highlights of Darwall's theory, the *Second-Person Standpoint* employs themes from Fichte's philosophy of right in order to strengthen the case for the inescapability of taking up the second-person standpoint of moral obligation. In his contribution for this special issue Darwall further develops his diagnosis that Fichte's thought offers in many respects a more promising, since more second-personal, foundation of morality than, for example, Kant's.

By now, the impact of Darwall's second-person standpoint theory has far transcended the confines of contemporary debates on moral obligation. Darwall has put to use the second-personal apparatus to critical engagements with Joseph Raz's theory of legal authority and Derek Parfit's

convergence arguments for his recent Triple Theory of moral wrongness. The constant theme that unifies all these diverse applications remains the one so impressively presented in *The Second-Person Standpoint*: without paying attention to the "interdefinable" and "irreducible" circle of (four) foundational second-personal concepts (valid demand, practical authority, second-personal reason, and accountability), neither superior epistemic status (Raz) nor the identification of optimific states of affairs (Parfit) are potent enough sources to generate anything close to the authority relationships that underlie the idea involved in obligating ourselves and one another. Given all of the above, it comes as no surprise that Darwall reserves his strongest sympathies for a specific ethical theory, namely contractualism. Our commitment to equal basic second-personal authority, that Darwall arrives at through his Fichtean rectification of the Kantian project, leads him to the endorsement of a contractualist paradigm in the spirit of broadly Rawls and Scanlon.

Grazer Philosophische Studien
90 (2014), 5–20.

WHY FICHTE'S SECOND-PERSONAL FOUNDATIONS CAN PROVIDE A MORE ADEQUATE ACCOUNT OF THE RELATION OF RIGHT THAN KANT'S

Stephen DARWALL
Yale University

Summary

The more foundational Part I of Fichte's *Foundations of Natural Right* was published in 1796, just before Kant's *Doctrine of Right* appeared in 1797. There are profound similarities in the ways Fichte and Kant treat matters of fundamental right as concerning reciprocal relations of freedom. However, I argue that there are also deep differences in the ways Fichte and Kant respectively ground natural right that give Fichte a better view. More specifically, I claim that the way Fichte brings a second-personal summons into the foundations of natural right as a call to the other freely "to determine itself in consequence of the summons" provides a potentially superior account to Kant's explication of right in terms of justified coercion. For Kant, "right and authorization to use coercion … mean one and the same thing." Like Kant, Fichte also recognizes a "right of coercion," but holds that it requires a justification that is downstream from any analysis of the concept of right itself. For Fichte, indeed, justified coercion presupposes a "common" and "reciprocal willing" of all persons that is part of a postulated "community" on which the normative force of natural right itself depends. Thus, whereas Kant takes a right against someone to consist in justification to coerce him, Fichte takes it to be something that is appropriately acknowledged in free compliance.

The more foundational Part I of Fichte's *Foundations of Natural Right* was published in 1796, just before Kant's *Doctrine of Right* appeared in 1797. There are profound similarities in the ways Fichte and Kant treat matters of fundamental right as concerning reciprocal relations of freedom. However, I shall argue that there are also deep differences in the ways Fichte and Kant respectively ground natural right that give Fichte a better view. More specifically, I shall claim that the way Fichte brings a second-personal summons into the foundations of natural right as a call to the other freely "to determine itself in consequence of the summons" (Fichte

2000, 35) provides a potentially superior account than Kant's explication of right in terms of justified coercion. For Kant, "right and authorization to use coercion … mean one and the same thing" (Kant 1996a, 6:232). Like Kant, Fichte also recognizes a "right of coercion," but holds that it requires a justification that is downstream from any analysis of the concept of right itself (Fichte 2000, 83, 88-92). For Fichte, indeed, justified coercion presupposes a "common" and "reciprocal willing" of all persons that is part of a postulated "community" on which the normative force of natural right itself depends. Thus, whereas Kant takes a right against someone to consist in justification to coerce him, Fichte takes it to be something that is appropriately acknowledged in free compliance.[1] I shall argue that Fichte's is a superior view of relations of right.

Before we begin, however, note how similarly Fichte and Kant view relations of right as concerned with maintaining the "external freedom" of beings who live alongside one another and who are capable of what Fichte calls "inner freedom" and Kant calls "internal", "law-giving", or "choice" (Fichte 2000, 10; Kant 1996a, 6:214).

Here is what they respectively say about the concept of right.

> Fichte: "The concept of right is the concept of the necessary relation of free beings to one another" (9).
>
> Kant: "Right therefore is the sum of the conditions under which the choice of one can be united with the choice of another in accordance with a universal law of freedom" (6:230).

And here is what they say about the "rule," "formula," or "law of right":

> Fichte: "The rule of right, limit your freedom by the concept of the freedom of all other persons with whom you come in contact" (10).
> Fichte: "the formula of right—limit your freedom so that others alongside you can also be free" (82).
>
> Kant: "The universal law of right, so act externally that the free use of your choice can coexist with the freedom of everyone in accordance with a universal law" (6:231).

1. Kant agrees that property rights can only be made determinate and conclusive in a civil, rightful condition that requires the united will of all.

That relating to others in a rightful way involves respecting and forbearing interference in others' "sphere" of freedom is common ground. Where Fichte and Kant diverge is in the way they respectively ground the relation of right and what they take the right to be one's own master within this sphere ultimately to amount to.

A Hohfeldian framework

It will help to provide a framework for the discussion to follow if we begin with some familiar distinctions between kinds of rights that are due to Hohfeld (1923).[2] According to Hohfeldian orthodoxy there are four analytically distinct things to which 'right' can refer: claims, privileges, powers, and immunities. Claim rights and privileges exhaust the kinds of "first-order" rights. Powers and immunities, on the other hand, are second-order phenomena involving capacities to affect, or resist, first-order rights.

The important first-order distinction is between claim rights and rights of privilege. A claim right is always to someone else's action or forbearance and entails a corresponding obligation that the person against whom the right is held owes to the right holder. Privileges, by contrast, entail no duties. To the contrary, a privilege to do something consists in the lack of any duty not to do it. Hobbes's "right of nature" that "each man hath, to use his own power, as he will himself, for the preservation of his own nature" is a privilege right in this sense (Hobbes 1994, 14.1). It is a "blameless liberty of using our own natural power and ability" (Hobbes 1983, I.14.6). Hobbes's right of nature thus entails no correlative duties. To the contrary, it consists in the right holder's being free of any duty that could constrain the right.

Powers and immunities are second-order rights in that they consist not in any particular position in the network of first-order claim rights, obligations, and privileges, but in conditions that can affect these. A power is an ability to change first-order rights and duties, as, for example, by promise or consent. And an immunity is an ability to resist attempts by others to change one's position in the first-order network. If someone is immune to prosecution, for example, then try as a prosecutor might, she cannot bring that person to trial.

2. For an excellent discussion, see Wenar 2011.

The important point for our purposes is that the fundamental right with which Fichte and Kant are concerned, the "rule" or "law" of right, is at least partly a claim right. Since it is a right of freedom of action, it may seem not to be a claim right in the sense that those created, for example, by promise or contract clearly are. It entails no obligation of others to perform any "positive" action, like returning a book or plowing a field. But unlike a mere privilege such as Hobbes's right of nature, it is something that others *violate* if they interfere with one's freedom of action (see, e.g., Fichte 2000, 101). If I am free to do whatever I judge prudent in preserving myself in the sense Hobbes has in mind, that just means that I do no wrong and do not wrong others in so acting. Such a privilege right is not something it is even possible to violate.

Fichte and Kant hold that there is a fundamental right to freedom of action within one's own sphere. This entails both a privilege of action within that sphere, but also a claim right that others not interfere by intruding outside of their sphere into one's own. Any such interference or intrusion is a violation of one's fundamental right. And that can be so only if the fundamental right includes a claim right held against others that it is possible (though not permissible) for others to violate in this way.

Kant is somewhat clearer on the fundamental rights including both the requisite privilege and claim right, but Fichte is committed to this also. Kant's "Universal Principle of Right" is a clear statement of the privilege: "Any act is right if it can coexist with everyone's freedom in accordance with a universal law" (6:230). In other words, any action within one's sphere of freedom, consistent with equal spheres for others, does not wrong others.

But Kant also makes clear that he is concerned with "the concept of right, insofar as it is related to an obligation corresponding to it," (6:230) that is, with rights that entail correlative (directed) obligations, hence with claim rights. Since the fundamental right of freedom is held against all persons, the correlative obligation not to interfere is owed by each person to every other.

Now Fichte says that "in the doctrine of right there is no talk of moral obligation" (Fichte 2000, 11). This has mainly to do with his view that any rights against others within communities must be derived from a person's "free arbitrary decision to live in community with others" and not from some fundamental moral obligation to do so. I will return to this later when we consider the role of the summons in Fichte's grounding of

natural right. We can set this aside for now, however, since even if there are no rights against particular others within one's community absent a "free, arbitrary decision to live with" them, what we are interested in is what Fichte is committed to when the right to freedom against others in fact holds. And it is hard to see how Fichte can avoid the conclusion that his fundamental right to freedom is a claim right if he is to think, as he clearly does, that it is something that others *violate* when they interfere with freedom of action.

The only kind of (first-order) right that can be violated is a claim right. And it is of the nature of claim rights to entail correlative obligations. It would seem, therefore, that both Kant and Fichte are committed to the idea that the fundamental right of freedom creates correlative obligations on all persons against whom the right is held.

The second-personal character of claim rights

Ultimately I want to discuss how Fichte's grounding of fundamental right in a second-personal summons gives him a superior view to Kant's. To help set up that discussion, however, let me first illustrate the irreducibly second-personal character of claim rights. Joel Feinberg made this fundamental point, though not in so many words, in his famous paper, "The Nature and Value of Rights." The eponymous nature of claim rights, according to Feinberg, is that they can be *claimed*. "It is claiming," Feinberg says, "that gives rights their special moral significance" (Feinberg 1980, 151).

Claiming is second personal in its nature. A claim must be addressable *to* the person on whom the claim is made. To have a claim right is, *inter alia*, to have the authority to claim it by addressing a claim to the person against whom the right is held. Nor is the authority to claim the only second-personal power or authority that is involved in a claim right. In having a claim right to non-interference, persons have, not just the authority to claim this right, but also, in many cases, at least, the authority to waive it, that is, to consent to actions of others that would otherwise constitute illegitimate interference. Consenting is also second-personal in its nature, like claiming; to be in force, consent must be given or addressed *to* someone who is, by virtue of it, released from an obligation he would otherwise have had.

Or consider what happens when a claim right is violated. At that point a new set of second-personal authorities entailed by the right kicks in.

When her right is violated, a victim has standing to relate to the person who has violated it in various ways that others may not. Whether to forgive a right's violation is distinctively up to the victim. Others don't have the standing to forgive that the victim has. I cannot forgive a right's violation of which you are the victim. Similarly, whether to seek apology is up to the victim in a way it isn't to others. Most obviously, an apology must be *to* the victim. But even seeking apology is the victim's prerogative. If your rights are violated, and I seek apology on your behalf, but without your authorization or, worse, against your will, then this may violate your rights further. The victim also has a distinctive authority to seek compensation. In American courts anyway, only victims have standing to bring tort actions.

In all these ways, to have a claim right is to have a set of second-personal powers or authorities *as the right holder*. I put this point in my work by saying that the right holder has an *individual authority*, as the very individual holding the right, to relate to those against whom the right is held in these second-personal ways. These second-personal features are reflected also in the correlative directed or relational obligations that the person against whom the right is held, the obligee, owes to the right holder. Just as the right holder has an individual authority to hold the obligee accountable to him for violations, so also is the obligee distinctively accountable *to the* right holder (Darwall 2012). That is what it is for the obligation to be owed *to* the right holder.

An adequate foundation for claim rights must account for their second-personal character. Before returning to Fichte and Kant, I want to note a final further aspect of claim rights' second personality, namely, the distinctively second-personal reasons for acting that derive from them. If I want to step on your foot, I have a reason not to do so in the fact that you have a right that I stay off your feet. To see this, notice that if you waive this right and consent to my stepping on your foot, this clearly affects the weight of reasons I have not to step on your foot. But if your waiving your right can cancel or lessen reasons not to step on your foot, then the fact that you had the right and had not yet waived it must have been a reason for me not to step on your foot in the first place.

Claim rights create what I call "second-personal reasons" for acting (Darwall 2006). What makes a reason second-personal is that it would not exist but for an authority to address the reason second personally in one or another of the ways we have been discussing. The existence of the reason does not depend on any actual address from you to me. You do

not have to claim the right in order to bring it into existence. Still, the right, and so the reason, would not exist but for your having the authority to claim it, hold me accountable for complying with it, and so on. And that is what makes the reasons for acting created by claim rights second-personal reasons.

Kant and right as an authority to coerce

Recall now Kant's claim that "right and authorization to use coercion … mean one and the same thing" (6:232). I take this to be Kant's analysis of a claim right. The right we each have that others not interfere with or "hinder" our freedom consists in our having justification to coerce them not to, to "hinder" their "hindrance," as Kant puts it.

But how does Kant derive this right? As I see it, he does so in two stages.[3] First, he argues that the "Universal Principle of Right" follows from fundamental features of the "concept of right," which he sums up in the formula: "Right is therefore the sum of the conditions under which the choice of one can be united with the choice of another in accordance with a universal law of freedom" (6:230). From this the Universal Principle of Right follows more or less directly: an "act is right if it can coexist with everyone's freedom in accordance with a universal law" (6:230). This, again, states a fundamental "privilege," any action that can coexist with everyone's freedom does no one a wrong; it is consistent with rightful relations to everyone.

Actions, however, that "hinder" others' freedom are not covered by this privilege. And Kant argues quite directly that just as "resistance that counteracts the hindering of an effect promotes this effect and is consistent with it," so also must "*hindering of a hindrance to freedom*" be "consistent with freedom in accordance with universal law" (6:231). It follows that the hindering of hindrances to freedom are rightful, covered by the fundamental privilege that is the Universal Principle of Right. But if that is so, then Kant evidently regards it as following from his definition of a (claim) right as justification to coerce that each person has a claim right that others not hinder his freedom since he can rightfully hinder their hindrance. Kant concludes: "Hence there is connected with right by the principle of

3. For an excellent discussion, see Ripstein 2009. For a discussion of Ripstein, see Darwall 2013b.

contradiction an authorization to coerce someone who infringes upon it" (6:231). The fundamental claim right follows from the concept of a claim right together with the Universal Principle of Right, which follows from the very idea of rightful relations between free persons.

But now note two consequences of Kant's analysis and derivation of the fundamental claim right to non-interference. The first is that no correlative *relational* obligation of others to respect the claim right follows. Of course, Kant believes that everyone is prohibited by the "universal law of right," not to mention by the Categorical Imperative, from violating the right to non-interference. But this is not a relational obligation *to* the right holder that is entailed by and correlative to the claim right he holds. Taken in itself, all his claim right involves is a justification to force others not to interfere with him.

On a Feinbergian analysis, a claim right consists in the second-personal standing to claim or demand of the obligee (the person against whom the right is held) that to which one has the right and to hold him accountable for providing it. Nothing like this can possibly follow from the claim right to non-interference as Kant analyzes and derives it.

The second consequence follows from the first. When someone stands on her right to non-interference and refuses consent, we ordinarily take it that she is, *inter alia*, asserting or implying that a distinctive reason exists for the person not to interfere, namely, that it would violate her right. But again, no such reason follows from the claim right to non-interference as Kant analyses and derives it. All that follows is that those against whom she has the right will not be wronged and cannot rightfully complain if she uses force to "hinder their hindrance." There may of course be reasons related to her right, or at least to her taking herself to have a right that will come into play, namely, that any attempt at interference will be resisted forcefully and indeed rightfully. But the fact that she has the right will not provide any reason against interference in itself. The point, again, is not that Kant has no way of generating a reason not to interfere with people's freedom. Both the Categorical Imperative and the universal law of right supply this. The problem is that neither of these follow from the fact that people have a claim right not to be interfered with and, consequently, that we are obligated *to them* not to do so.[4]

4. At least, this does not follow in the sense we normally have in mind when we speak of obligations. In the sense Kant uses 'obligation' ('*Verbindlichkeit*'), however, it might be argued to follow in the following way. Kant distinguishes between moral and practical "necessity," on the one hand, and what he calls "necessitation," on the other (6: 222). The former is a thoroughly

My claim is that Fichte's derivation of natural right provides materials to resist these two consequences of Kant's doctrine of right and that it can consequently yield a more adequate account of fundamental claim rights. I am not saying, I should emphasize, that Fichte would necessarily have agreed with this claim. As I mentioned earlier, Fichte resists talk of obligation, draws a sharp line between morality and right, and derives claim rights against particular others within one's community from their free "arbitrary" decision to live in community with the agent, "not through any obligation" (15). Nevertheless, Fichte's strategy is clearly to "deduce" the concept of right as a "condition of self-consciousness" and thereby to show that it is "an original concept of pure reason" (9). And there is no question that he takes a reciprocally recognizing, and thus second-personal, summons from a free rational being to the agent (as a free rational being) to be necessary for self-consciousness. My claim is that the best philosophical interpretation of Fichte's thesis that fundamental claim rights can be grounded in a presupposed summons *a priori* entails a presupposition of equal basic second-personal authority that can ground the second-personal nature of fundamental claim rights and the second-personal reasons for acting to which they give rise.

Consider, first, how Fichte argues for the necessity of a summons for self-consciousness as a free agent.[5] "In acting," Fichte says, "the rational

---

normative fact whereas the latter is ultimately *psychological*: an imperfect rational being's determining herself to do what the law objectively required. Strictly speaking, then, only beings for whom compliance with the objective law is "subjectively contingent" can be under obligation (4: 413). "Obligation involves not merely practical necessity (such as a law in general asserts) but also *necessitation*" (6:223). Although all rational agents, including "holy beings," are under the moral law, only imperfect rational agents are under obligations or "imperatives;" and they determine or "constraint" themselves to follow the law through addressing reasons' commands to themselves and following them. This means that obligation is ultimately a psychological phenomenon for Kant. Moreover, Kant distinguishes between "internal" and "external constraint." We imperfect rational agents act *ethically* through *self-constraint*, when we give ourselves commands of reason and follow them. By contrast, the philosophy of right is concerned with *external* constraint of action, both by others and, in a civil condition, by the state. Rights for Kant involve authorization to constrain others' conduct. But that implies that they can "necessitate," and, in that sense, "obligate" others' not to violate their rights, by hindering their hindrance. On Kant's use of 'obligation,' therefore, claim rights do imply correlative obligations to right holders. I hope it is obvious, however, how Kant's ultimately psychological sense of 'obligation' expresses something different from the distinctive normative items we normally take obligations to be.

5. In what follows, I draw from Darwall 2013c.

being does not become conscious of its acting; for *it itself* is *its acting* and nothing else" (4). The agent is conscious of its "object," "of what emerges for it in this acting" (4f.), not of itself as a self-determining will. Fichte's idea, as I interpret it, is that the first-person perspective of unsummoned agency is focused on various states of the world the agent can bring about by her actions, the relative desirability of these, what it would take to bring them about, their costs and benefits, and so on. The objects of unsummoned practical thought are the alternative actions before the agent, and the practical question is which to choose.

The agent's focus is, as it were, outward, on the objects before her, not inward on her own agency. For an unsummoned agent, thoughts of her own free agency can have practical relevance only insofar as they relate to the objects before her—e.g., the desirability of bringing some state of the world about, what it would take to do that, and so on. Any attempt to focus on her own free agency as such risks a futile attempt to observe her own agency.

For a subject to gain self-consciousness of herself as a free agent, the "object" of the subject's consciousness must be "synthetically unified" with the "subject's efficacy." The object of consciousness must be "nothing other than the subject's efficacy" itself, so that "the two are the same" (31). Fichte's transcendental claim is that this can be achieved only *second*-personally, by a "summons" from one rational agent to another "calling upon it to resolve to exercise its efficacy" (31). The summons is to the other *as an agent*, so in being aware of it, the other is aware of herself as thus regarded. So far, however, this might be no different from a *third*-personal, observer's awareness. That the other sees her as an agent, or even that she so sees herself, is but another aspect of the way things are anyhow.

What makes all the difference is that a summons addresses the agent *second*-personally; therefore in taking it up the agent *per force* relates-to-the-other-relating-to-her-as-an-agent. She operates within a second-personal relationship in which each reciprocally recognizes the other as a "you" to whom each is a "you" in return. The presuppositions of intelligible second-personal deliberative thought simply require that the agent *deliberate under the assumption* that she (and her co-deliberator) are both free agents. This gives her a practical awareness of her agency that is irreducible to any consciousness she has of herself as part of a causal order. *She simultaneously "posits" herself and the other as free and rational agents within her own deliberation*, from the practical point of view (9). She grasps herself "in this identity of acting and being acted upon" (23, see also 40).

The most perspicuous interpretation of Fichte's idea is in terms of second-personal reasons. *A summons is any address of second-personal reasons to another agent.* Since it addresses the other *as agent*, a summons necessarily involves the giving of reasons. "The rational being's activity is by no means to be determined and necessitated by the summons in the way that … an effect is necessitated by its cause; rather the rational being is to determine itself in consequence of the summons" (35). A summons, whether a claim, demand, or request, attempts to give another agent reasons by which she can freely determine herself, reasons that are grounded in a presupposed authority to summon the agent to act in some way, even if only to figure out for himself what to do.

As a contrast, consider simply pointing out to someone that there is reason for her to act owing to the desirability of what she would thereby bring about. So far, you only summon her to *believe* that this is what she has reason to do, but not actually *to do* it. You summon her as a being that can form beliefs about what to do, not directly as an agent. For her to take up your summons she must deliberate, not about what to do, but about what to believe she should do, on the assumption that she can freely form beliefs in response to epistemic reasons.

But now consider a case in which you summon her to act, whether to do something: say to bring about a certain desirable state of affairs, to provide you something to which you have a legitimate claim, or even just to make up her own mind what to do without any further direction from you. In all these cases, including both the first *and* the last, you make a claim on her will. You purport to give her a reason to do something, even if only freely to decide what to do, that is additional to any reasons that she would have had independently of the standing you presuppose to summon her. If she takes up your address (which she cannot avoid doing if it is common between the two of you that she has listened and heard), then she reflects back a reciprocal address (as someone who, like you, has the standing to address second-personal reasons as well). Even a bare request addresses a second-personal reason that is additional to any non-second-personal reasons that might stand behind it, since it presupposes the normative standing to make the request.

In *The Second-Person Standpoint*, I argue that any putative address of second-personal reasons is committed to both a shared second-personal competence to choose freely to act on such reasons and a shared basic second-personal authority to make claims and demands of one another from which such reasons flow (Darwall 2006). In my view, it is this shared

second-personal authority that underlies fundamental claim rights. Fichte's position is that the concept of right can be deduced as a "condition of self-consciousness" and that a second-personal summons is a condition of thought. The question is what is the relation between the summons and the concept of right.

In my view, Fichte should hold that the connection is direct. Any summons that purports to give an addressed agent a second-personal reason for acting, even just to make up his own mind and act, presupposes the capacity of addresser and addressee freely to determine themselves by reasons that are grounded in a shared second-personal authority it simultaneously presupposes. Fichte's own position is, however, somewhat equivocal. On the one hand, he says things that seem to deny this possibility. But other things he says seem to require it.

For Fichte, the "concept of right," again, "is the concept of the necessary relation of free beings to one another" (9). The "*complete object* of the concept of right [is] … *a community of free beings as such*" (10). In a community of right each recognizes others' right of "external freedom" through "inner freedom." Fichte denies that there is any obligation *of right* to enter into a community of right. "The thought and task of such a community is arbitrary," Fichte says, so the concept of a community of right is only "technical-practical," that is, "*if* someone wanted to establish" "a community among free beings," it would have to be done "in accordance with the concept of right" (10). This would clearly seem to rule out the line of thought I am suggesting.

Fichte allows that "it has indeed been shown that, if a rational being is to come to self-consciousness—and hence is to become a rational being" this can only be done through reciprocally recognizing second-personal rational influence that itself involves recognizing the other's sphere of external freedom through internal freedom. "But," he says, "that even after self-consciousness has been posited, rational beings must continue to influence the subject of self-consciousness in a rational manner, is not thereby posited" (81). The "enduring existence" of a community of free beings related by right is thus an "arbitrary," rationally optional "postulate" that an agent may or may not adopt "of his own free choice" (81). Although there may be "an obligation to will this" "within the sphere of morality," "in a theory of natural right" all that can be said is that *if* an agent wills to be a member of a community of free beings, then he is bound to them by relations of right (81).

On this "voluntarist" interpretation, it takes an individual's voluntary participation in a "reciprocal declaration" to be obligated to another by

the principle of right (15). Each obliges the other and himself by an actual reciprocal recognition (44). Moreover, both the recognition and the obligation it gives rise to are reciprocally conditional. "This manner of treatment is conditioned by the first's treatment of the other; and the first's treatment of the other is conditioned by the other's treatment and knowledge of the first" (42).

Nonetheless, there are reasons for thinking that Fichte cannot hold to a voluntaristic interpretation if he is to maintain that the conditions for self-awareness are sufficient to validate the principle of right. Fichte's official position, again, is that it is only a voluntary "reciprocal declaration" of recognition with a specific individual that obligates the agent to recognize that individual's sphere of freedom and to limit his own. However, even here, Fichte says that although one cannot complain that another does one an injury in refusing recognition, the "doctrine of right" nonetheless warrants the claim that the other "must then remove himself from all human community" (12). This suggests, first, that Fichte believes the only way an individual can avoid the obligations of the principle of right is to avoid other people altogether. But, second, it is not clear why this should be so on voluntarist assumptions. Why wouldn't there simply arise various *communities* of right, that is, associations within which individuals are obligated by the principle of right, with no obligations of right to outsiders? It is hard to see how a voluntarist interpretation can avoid this consequence.

Other things Fichte says fit no better with a voluntarist interpretation. First, Fichte asserts that agents demand continued recognition of themselves and their freedom "for all the future" when they reciprocally recognize one another (48). But again, why should this be so on a voluntarist interpretation? It would seem that individuals would be as free voluntarily to obligate one another for a temporally limited period as to do so indefinitely. Of course, if, as I have been suggesting, reciprocal recognizing individuals are committed to the claims rational persons have *as such* to address second-personal reasons, not as a matter of voluntary agreement, but as a presupposition of their intelligibly addressing claims to one another second-personally at all, then such a demand would be expected.

Fichte also frequently says that reciprocally recognizing agents recognize one another *as* rational beings, and that they are thereby committed to treating one another *as* rational beings or "persons" (e.g., 42, 43). But these claims presuppose that there are ways mistreating rational beings *as*

such and, therefore, that failing to recognize a rational being is not simply forbearing to make a voluntary commitment one is free not to make. They presuppose that rational being is itself a normative standing, that there are ways of respecting or mistreating them just by their nature as rational beings capable of reciprocal recognition (second-personally competent), and that, therefore, one is not free not to recognize and respect them. Again, this makes perfect sense on a "presuppositional" interpretation, but not on a voluntarist one.

Finally, the most significant problem with the voluntarist interpretation is that, unless we assume background relations of right and an obligation to keep voluntarily made commitments, a voluntarist interpretation is powerless to explain how reciprocal recognition can give rise to any obligation to respect spheres of freedom. Fichte clearly assumes that individuals have warranted claims against each other if they violate the other's sphere of freedom once reciprocal recognition has transpired. If I have conformed to the law we both committed ourselves to in reciprocally recognizing one another, and my co-respondent subsequently violates that law, then I am in a position to charge him with a violation of my right.

> I … appeal to a *law* that is valid for us both, and apply that law to the present case. I thus posit myself as judge, i.e. as his superior. . . . But, insofar as I appeal to that common law in my opposition to him, I invite him to be a judge along with me; and I demand that in this case he must find my action against him consistent and must approve of it, compelled by the laws of thought (47).

But what gives any "law" we voluntarily commit ourselves to normative force? The fact that we committed ourselves to it, as if adopting it together? Unless we assume that we each already have the normative standing to obligate ourselves through our reciprocal commitments, no reciprocal willing can yield any obligating law.

Fichte says that once someone has (arbitrarily, reciprocally) willed to live in community with others—has accepted this "hypothetical" "postulate"—she "must also necessarily will the law," "thus the law has hypothetical validity" (82). Fichte's terming the concept of right a "technical-practical" concept encourages us to read him as saying that since relations of right are necessary to any community of free beings, any agent who wills the latter must will the former, as Kant says about hypothetical imperatives, "insofar as reason has decisive influence on his actions" (Kant 1996b, 4:417). The problem is that this just assumes that rational agents *can* impose obligations

on themselves, but they cannot without the requisite authority and that authority cannot *itself* be explained by its being necessary to something they arbitrarily will. From the fact that relations of right are necessary to something I will nothing can possibly follow that could actually make me subject to them. And if we assume that you and I have the authority or power to obligate ourselves by our reciprocal willing, then we unavoidably assume that we stand in relations of right already.

My alternative suggestion is that we interpret Fichte as saying that whenever you and I enter into a reciprocally recognizing, second-personal relation of summoned and summoner we are committed thereby to assuming that we each have the standing as beings who are capable of rational second-personality to be sources of second-personal reasons for one another. In seeking treatment from one another as rational second persons, we do not *endow* one another with this dignity or standing; we presuppose that we each independently have it.

Interpreting Fichte in this way builds basic second-personal authority into the transcendental second-personal condition of self-consciousness in a way that allows us validly to deduce the concept of right. Moreover, it can account for the second-personal aspects of fundamental claim rights, and the second-personal reasons for acting to which these give rise, in a way that Kant's theory cannot.

References

Darwall, Stephen 2006: *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

— 2012: "Bipolar Obligation". In: Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics*, *v. vii*. Oxford: Oxford University Press, 333–367. Also in: Darwall 2013a.

— 2013a: *Morality, Authority, and Law: Essays in Second-Personal Ethics I*. Oxford: Oxford University Press.

— 2013b: "Forcing Freedom. Critical Review of Arthur Ripstein's *Force and Freedom*". *Legal Theory* 19, 89–99.

— 2013c: „Fichte and the Second-Person Standpoint". In *Honor, History, and Relationship: Essays in Second-Personal Ethics II*. Oxford: Oxford University Press, 222–246.

Feinberg, Joel 1980: "The Nature and Value of Rights". In: Joel Feinberg (ed.), *Rights, Justice, and the Bounds of Liberty*. Princeton, NJ: Princeton University Press, 159–184.

Fichte, Johann Gottlieb 2000: *Foundations of Natural Right.* Cambridge: Cambridge University Press. Translation by Michael Bauer.

Hobbes, Thomas 1983: *De Cive, the English Version, Entitled in the First Edition, Philosophical Rudiments Concerning Government and Society*. Oxford: Clarendon Press. Translation by Howard Warrender. (References are to book, chapter, and paragraph number.)

— 1994: *Leviathan*. Indianapolis, IN: Hackett Publishing Co., Inc. References are to chapter and paragraph number.

Hohfeld, Wesley Newcomb 1923: *Fundamental Legal Conceptions as Applied in Judicial Reasoning and other Legal Essays.* New Haven, CT: Yale University Press.

Kant, Immanuel 1996a: "The Metaphysics of Morals. Part I: Metaphysical Principles of the Doctrine of Right". In: Kant, *Practical Philosophy*. Cambridge, MA: Cambridge University Press. Translation by Mary Gregor. (References are to the Preussische Akademie edition and, second, to the page numbers of the Cambridge volume.)

— 1996b: "Groundwork of the Metaphysics of Morals". In: Kant, *Practical Philosophy.* Cambridge, MA: Cambridge University Press. Translation by Mary Gregor. References are to the Preussische Akademie edition and, second, to the page numbers of the Cambridge volume.

Ripstein, Arthur 2009: *Force and Freedom: Kant's Legal and Political Philosophy*. Cambridge, MA: Harvard University Press.

Wenar, Leif 2011: "Rights". In: Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. URL= <http://plato.stanford.edu/entries/rights/#2.1>.

Grazer Philosophische Studien
90 (2014), 21–34.

SECOND-PERSONAL REASON-GIVING

Fabienne PETER
University of Warwick

Summary
David Enoch has recently objected to Stephen Darwall's account of second-personal reason-giving that the phenomena that Darwall focuses on can be fully explained without resorting to second-personal reasons. In this paper, I shall argue, against Enoch, that second-personal reason-giving matters. My account of second-personal reason-giving differs from Darwall's, however, as it accepts that some of the phenomena Darwall focuses on can be reduced to the more standard form of reason-giving.

Introduction

The key insight that Stephen Darwall develops in *The Second-Person Standpoint* is that our interactions with others can be of fundamental normative significance. It is nothing new, of course, that our interactions with others may causally affect us. If you are in a bad mood, I might not enjoy being around you as much as I normally do. If you tell me about a recent trip you made, I might form a desire to visit those places too. And if you tell me that you have just read that it will rain again this afternoon, this might cause me to form a corresponding belief. But we can establish those causal effects without establishing anything about the normative significance of these interactions. The normative significance of our interactions with others depends on how they relate to our (normative) reasons for action or for attitudes such as beliefs, desires, etc. To keep things simple, I shall focus here on reasons for actions—the case of practical normativity.

It is also fairly uncontroversial to say that our interactions with others may be normatively significant in the sense that they trigger certain reasons. For example, if I act dismissively towards you, this may give you a reason to express blame or to avoid me. The reason to blame or to avoid disrespectful people is not, I take it, created by my action. It has been there all along.

But my action may trigger this reason, such that your normative situation is now no longer the same as it was before I interacted with you in this way.

Darwall has drawn our attention to the possibility that our interactions with others can be normatively significant in a more fundamental way: they may not just trigger reasons that have been there all along; they may also create reasons that would not exist without this interaction. Second-personal reasons are of this kind.

One of Darwall's favourite examples is the following. Suppose I accidentally stepped on your toe and you claim that I should remove my foot from on top of yours. Darwall argues that this gives me a reason to remove my foot that was not there before you claimed that I should, a reason that is different from other reasons I might have to remove my foot, say reasons grounded in sympathy. It is a second-personal reason that depends for its existence on properties of our relationship. It depends on whether you have the authority to make such claims and, vice versa, on whether I am accountable to you in this respect. Such second-personal reason-giving, Darwall claims, plays a role in requests, commands, promises, and similar practical scenarios and its role is essential for explaining moral obligations.

David Enoch (Enoch 2011, 2014) has recently challenged Darwall's account of second-personal reason-giving. Enoch argues that Darwall's account is not only metaphysically fishy—it appears that reasons are created out of thin air—it is also unnecessary: we can account for the kind of normative phenomenon that Darwall focuses on—Enoch calls it "robust" reason-giving—in terms of triggering reason-giving.

My aim in this paper is to defend the normative significance of second-personal reason-giving against Enoch's objections. My defense will, however, offer an interpretation of second-personal reason-giving that differs somewhat from Darwall's own interpretation and that assigns it a different place in our normative geography, as it accepts that some of the phenomena Darwall focuses on can be reduced to the more standard form of reason-giving.

Darwall on second-personal reasons

As Darwall influentially put it, second-personal reasons are reasons whose

> validity depends on presupposed authority and accountability relations between persons and, therefore, on the possibility of the reason's being addressed person-to-person. (Darwall 2006, 8)

Second-personal reason-giving is from within what Darwall calls the second-person standpoint. The second-person standpoint is a web of four concepts—reason, claim, practical authority, and accountability—where each entails the others. Someone's valid claim gives someone else a reason to act accordingly. To use Darwall's own example again, your claim that I move my foot from on top of yours gives me a reason to do so. And it does so, not because your claim triggers my sympathy, but because we mutually acknowledge our authority and accept our accountability to each other in cases like that—we each accept that you have a right to make claims of this sort on me and, vice versa, that I have a right to make this sort of claim on you. In other words, what makes your claim valid is that I accept my accountability to you in this regard or, which is the same thing, that I accept your authority to make such claims. Without that presupposed relationship of authority and accountability, your claim would not be valid and would thus not give me a reason to act accordingly.

The web between the four concepts characterizes the relationship between moral agents and exhaustively explains, according to Darwall, the normative grip of each of the components. What is distinctive about the second-person standpoint is that it locates a source of normativity in the relationship between moral agents—not in individuals as such (first-person standpoint) and not outside of their relationship (third-person standpoint).

The divisions that Darwall draws between the second- and third-person standpoints, on the one hand, and, on the other, between the second- and first-person standpoints are not equally sharp. The significance of the second-person standpoint in morality is linked to a denial of a third-personal, or fact-relative, source of moral normativity. So the distinction between the second- and the third-personal standpoint in ethics is, on Darwall's account, a sharp one. As Darwall explains (Darwall 2006, 8): second-personal reasons "simply wouldn't exist but for their role in second-personal address". As I interpret it, the key claim that *The Second-Person Standpoint* makes in this regard is a constructivist one: second-personal reasons are agent-relative in the fundamental (or metaphysical) sense that Korsgaard (Korsgaard 1996) has identified: they are reasons that depend on our agency for their existence.

Because second-personal reasons depend on our agency, first-personal considerations are part of the second-person standpoint (Pauer-Studer 2010). But this does not imply that the second-person standpoint reduces to the first-person standpoint. Instead, the second-person standpoint serves

to qualify first-personal considerations; it is necessary to establish what we can validly claim from each other.

Enoch on robust reason-giving

Enoch accepts that there seems to be something special about reason-giving in contexts such as requests or commands that a theory of practical normativity needs to be able to explain. But he rejects the explanation that Darwall's account of second-personal reasons offers.

Enoch (Enoch 2011, 2014) distinguishes between three different senses in which interactions with others might give you reasons for action. The first is the epistemic sense. Epistemic reason-giving occurs, for example, if you draw my attention to a relevant practical consideration I had overlooked. To illustrate, you might point out to me that it is my grandmother's birthday tomorrow. This gives me a reason to call my grandmother. But the reason for action was there all along, independently of our interaction.

Second, there is the reason-triggering sense that I have already mentioned. Triggering reason-giving occurs if the interaction with others manipulates the non-normative circumstances such that a latent reason becomes active. Again, the reason is there all along; it is not created by the interaction. Enoch gives the example of someone setting a foot on the road, thus activating the reason for drivers "to-stop-should-a-pedestrian-start-crossing" (Enoch 2011, 4).

Enoch calls the third sense "robust reason-giving". That is the reason-giving that occurs in contexts such as request, orders, or similar practical scenarios. For example, if a police officer tells me to stop my car, this will give me a reason to do so (and I probably ought to do it). Enoch uses the example of a request: if I ask you to read a draft of my paper, this gives you a reason to do so.

Robust reason-giving is clearly important in our practical lives, Enoch agrees. But how should we account for it? In terms of second-personal reason-giving, as Darwall claims? Does the power of requests or orders to generate reasons depend on the right of those making such demands and the accountability of the addressees? Enoch rejects that move and defends a reducibility claim instead: robust reason-giving is nothing but a particular instance of triggering reason-giving. While it appears as if there was something normatively special—perhaps second-personal—about robust reason-giving, properly understood this turns out to be misleading:

> when I ask you to read my paper, presumably there is this general reason (to do as I ask, within limits, in a certain context, etc.), one that I presumably succeed in triggering by making the request. It is in this way, then, that the suggested account of robust reason-giving is a particular instance (but an importantly unique one) of triggering reason-giving. (Enoch 2011, 16)

Enoch's argument for the reducibility claim invokes the distinction between a wide scope and a narrow scope reading of the conditional reason that I have to do as you ask me to (Enoch 2011, 10).

> Wide scope: You have a reason to (read the draft if I ask you to read it).
> Narrow scope: If I ask you to read the draft, you have a reason to read it.

The wide scope reading shows how the reason-giving in question can be interpreted as a version of triggering reason-giving. The narrow scope reading, by contrast, does not entail triggering reason-giving. If we only focus on the narrow scope reading, it appears as if robust reason-giving is distinct from triggering reason-giving. And if that were the correct account of robust reason-giving, it would be compatible, for example, with second-personal reason-giving.

Enoch argues, however, that while the narrow scope reading gives a plausible description of the practical phenomenon, it is only plausible in conjunction with a wide scope reading. The thought is that the wide scope reading explains the truth of the narrow scope reading and there is no alternative, better explanation of its truth. Since the wide scope reading characterizes triggering reason-giving, we have established the reducibility of robust reason-giving to triggering reason-giving. And if the reducibility claim is true, there is no need to resort to second-personal reasons to account for the normative phenomenon of robust reason-giving.

In addition, if the reducibility claim is true, then there is no further source of practical normativity—contrary to the claim that there are normatively distinctive, second-personal reasons. All normativity stems from the reasons that there are, i.e. from reasons that are not constructed and agent-relative in the fundamental sense that Darwall describes.

Responding to normative practical uncertainties

Is the reducibility claim correct? And do the implications that Enoch draws from it for the prospects of Darwall's account of second-personal reason-

giving hold? I want to grant to Enoch that some instances of robust reason-giving, including some that look like instances that Darwall's account of second-personal reasons aims to cover, do indeed reduce to triggering reason-giving. Even Darwall's favored foot example may be a case in point. But granting such a weak interpretation of the reducibility claim—one that says that *some* instances of robust reason-giving reduce to triggering reason-giving—still leaves room for an account of second-personal reason-giving that is normatively significant. I reject the strong interpretation of Enoch's reducibility claim, according to which *all* instances of robust reason-giving reduce to a form of triggering reason-giving.

The account I shall offer diverges in an important respect from Darwall's original account of second-personal reasons, however. According to Darwall, second-personal reasons are fundamental for moral obligations. I propose to treat them as residual instead: second-personal reason-giving matters in contexts where third-personal reasons underdetermine how we ought to act. Specifically, I have in mind epistemic underdetermination. My argument, but not Darwall's, starts from the uncertainties that affect our practical deliberation, deliberation about how we ought to act. These uncertainties are particularly salient in interactional contexts and they can explain why second-personal reason-giving matters.

Practical uncertainties may be of a non-normative kind. We often do not have a completely accurate picture of the circumstances we are in and cannot anticipate with certainty all the consequences of our actions. For example, if I leave the house now and make my way to the station without delay, can I still catch the 2:30pm train to Oxford? There is a certain probability that I will catch the train and a certain probability that I will not. This is not the kind of uncertainty I am concerned with here. The second kind of practical uncertainties is normative. It concerns the source of our normative beliefs and, as a result, the epistemic status of those beliefs: how can I know the difference between a true normative belief that correctly represents the normative reasons that apply and a belief that appears normative to me but that is simply the product of evolution or some other natural process? For example, is my inclination to offer help to someone supported by a correct belief that I have reason to help this person or is it merely a conditioned response that lacks warrant?

In his recent discussion of Derek Parfit's *On What Matters*, Darwall (Darwall 2014, 91) argues that in light of this epistemological challenge as well as of our tendency to disagree in normative matters, doubt arises as to whether there are any normative facts at all. Parfit is aware of the

problem, but defends the claim that practical normativity derives from normative facts. His argument hinges on a convergence claim. Parfit (2011) argues that our normative uncertainties and disagreements are merely superficial; underlying those disagreements is a substantive convergence among the three main normative theories—consequentialism, Kantian contractualism, and Scanlonian contractualism. This convergence is best explained, Parfit argues, as driven by the normative facts that the different theories attempt to track. If the convergence claim were true, we could identify correct beliefs about normative facts in the area of convergence among those theories and thus eliminate uncertainties. Darwall rejects Parfit's solution, however, and argues that the convergence claim is false (Darwall 2014, 99ff).

If the convergence claim is false, and I agree with Darwall that it is, normative practical uncertainties (and the resulting disagreements) are back. But the correct response to normative practical uncertainties, I believe, is not to doubt the existence of normative facts altogether. The correct response, instead, is via a fuller development of practical justification under circumstances of uncertainty. When trying to establish how we ought to act, we often do not have knowledge of the reasons that apply. Vice versa, we may act on what we take to be reasons that apply even though our beliefs are false and hence we may not do what we have reason to do. Parfit is, of course, aware of this predicament. To capture the subjective element in practical deliberation, Parfit draws a distinction between "apparent" reasons and "real" reasons (Parfit 2011, 35).

What is the normative status of such apparent reasons? Parfit answers this question by drawing a distinction between responding to reasons and being practically rational. Acting on apparent reasons is all that practical rationality requires, he claims:

> Our desires and acts are rational when they causally depend in the right way on beliefs whose truths would give us sufficient reasons to have these desires, and to act in these ways. (Parfit 2011, 112)

Acting on apparent reasons may thus be permissible in the sense of being rational. But apparent reasons do not have normative force, only real reasons do (Parfit 2011, 35).

Parfit further develops his view by distinguishing between different senses of ought (Parfit 2011, 150f). He associates full-fledged normativity with a fact-relative sense of ought—what we ought to do is determined by the real reasons that apply. The ought of practical rationality is belief-

relative.[1] Finally, there is also an evidence-relative sense of ought, but that one plays no role in Parfit's theory of practical normativity.

While I think that Parfit moves in the right direction here by recognizing how we often have to act without knowledge of the relevant normative facts, there is something puzzling about the idea that there are oughts that do not entail a normative reason. The problem with this way of understanding apparent reasons is that it is either incompatible with the claim that normativity is about reasons or with the claim that what we ought to do is normative (Kiesewetter 2012).

Part of this problem is Parfit's way of characterizing practical rationality. This characterization is controversial, mainly because it focuses only on the belief-relative sense of ought and not the evidence-relative sense. I do not find it plausible. But since I am concerned here with the question of how we should understand normative reasons, this issue is tangential to the topic of my paper and so I want to bracket it.

Parfit's understanding of normative reasons as giving rise to a fact-relative sense of ought is another part of the problem. While Parfit can allow for the possibility that the belief- and even the evidence-relative sense of ought influence our practical deliberation, they only do so at the level of apparent reasons. Normative reasons set a standard for the success of practical deliberation, but practical deliberation itself cannot give rise to normative reasons.

I do not think that this is right; this understanding of normative reasons rests on a truncated conception of our practical agency. In light of the practical uncertainties that surround practical deliberation, we need to introduce a further distinction, beyond the distinction between apparent reasons and real reasons. The distinction is between apparent reasons and constructed reasons. The thought is the following. There is an intuitive difference between a consideration that we mistakenly take to be a normative reason—because we hold a false belief—and a consideration that we take to have normative force in the absence of knowledge about the real reasons that apply to us. In Parfit's use of the term, both kinds of considerations fall under the category of apparent reasons and both may give rise to oughts, even though neither has the normative force of real reasons. But I do not see why we should elevate our simple mistakes in this way. Sure, we may sometimes be excused

1. "[W]e ought rationally to act in some way when this act is what we ought practically to do in the belief-relative or normative-belief-relative sense" (Parfit 2011, 163).

from making mistakes. But to be excused from making a mistake is not the same thing as acting as one ought to, in any sense of ought. By contrast, it seems plausible to me that there are at least some responses to the uncertainty we face about which of our normative beliefs are true that can, as such, give rise to an ought. They cannot, of course, give rise to an ought in the fact-relative sense. But they nevertheless have independent normative force.

Based on this distinction we get three categories: (1) apparent reasons that are based on true beliefs about normative reasons—these are also real reasons, (2) apparent reasons that are given by mistaken normative beliefs and that do not have independent normative force, and (3) reasons that are constituted by a commendable response to the uncertainty we face about real reasons and that have independent normative force. I want to call the reasons of the third category constructed reasons. Constructed reasons are agent-relative.

A lot more would have to be said, quite generally, about the properties of a commendable response to normative practical uncertainties, but I cannot do this here. All I can do is discuss the rough idea with regard to the issue under debate, namely whether there are second-personal reasons or whether Enoch's reducibility claim about robust reason-giving is true.

Defending second-personal reason-giving

Enoch's alternative to Darwall's account hinges on the assumption that only what I have called, following Parfit, real reasons can have normative force in the sense of having the capacity to shape how we ought to act. As long as the relevant facts obtain, these reasons have normative force. Whether the relevant facts obtain depends on reason-triggering factors. To go back to Enoch's earlier example, once the pedestrian signals his intent to cross the road, the driver has reason to stop. The signal has triggered the relevant reason. Epistemic reason-giving, i.e. citing reasons for belief about the relevant facts, can help us form correct beliefs about how we ought to act. If I am a passenger in your car, I might shout, "stop—didn't you see the pedestrian?" to make you aware of what you have—real—reason to do.

On Enoch's account, robust reason-giving also needs to be explained in terms of real reasons. The interaction in those contexts in which robust

reason-giving occurs is such that it triggers a latent real reason for action. For example, if my request that you read a draft of my paper truly triggers the conditional (read the draft if I ask you to read it), then my request succeeds in giving you a—real—reason to read it.

Importantly, on Enoch's account, you have this reason independently of whether you are aware of the conditional or regard it as true. More generally, your attitudes towards the request are not part of the picture. Of course, my successful request does not imply that you now ought to read my draft—you may have much better things to do. But my successful request triggers a reason for action that changes your normative circumstances.

I want to grant Enoch that his account works for cases like the one just discussed. But I do not think that it works in all cases. First, consider what happens if the request is outlandish. Suppose I ask you to write my paper for me. I actually do not believe that you have a reason to write my paper for me and, I assume, neither do you. So we can agree that the conditional (write my paper if I ask you to write it) is false. If we are right, Enoch's account implies that my request does not trigger a reason for action for you and your normative circumstances are thus unchanged. Note that this implication is due to the fact that the conditional is false, not due to the fact that we agree that it is false.

Now consider the next case: what happens if we are uncertain about the truth of the relevant conditional? As it happens, I believe that this is not unusual; quite to the contrary, such uncertainties are ubiquitous and have a deep impact on our practical deliberation. How should we conduct our relationships with others? How can we establish whether our cherished goals deserve the attention we are inclined to give them? Should we obey an authority's directive that appears unjust or otherwise wrong? These are just a few examples of the kind of normative uncertainties that we often encounter. Suppose I am asking you to finish my paper for me because I find myself under unusual pressure, for example because I am currently suffering from some illness but my career hinges on that paper being completed. So the relevant conditional becomes something like (write my paper if unusual pressures force me to ask you to write it). This conditional (or a close cousin) might be true or it might not; I do not know. Suppose you do not know either.

On Enoch's account, all that we can say about a case like this is the following: if the conditional is true, then you have a reason to read my paper; if the conditional is not true, then you do not. If practical normative uncertainties are as ubiquitous as I believe they are, this response is

somewhat unhelpful. I also think it is false. When we face uncertainties about the truth of the relevant conditionals, those normative uncertainties do not necessarily give rise to normative underdetermination. Qua practical agents, we are able to respond to those uncertainties and, in so doing, fill the gap that is left by our lack of knowledge about the real reasons that apply to us. One way to put this is that responding to reasons is not the only function of practical deliberation. Deciding what to do when we do not know what reasons apply is another. With regard to the latter, the important thing to notice is this: some of these decisions will be better than others. But since we have already established that normative reasons underdetermine what we ought to do, we need a normative standard other than real reasons to make sense of this idea.

I want to propose that we think of second-personal reasons as setting such an alternative normative standard. Recall that Darwall's understanding of moral normativity in terms of second-personal reasons is in contrast to a fact-relative, third-personal understanding. What I am proposing is different: second-personal reason-giving does not replace third-personal reason-giving, not even in the moral domain. Instead, the two forms of reason-giving complement each other, given circumstances of normative practical uncertainty. In short, my basic idea is this: when we face uncertainties about the truth of a conditional, there is a second route to establishing its validity. This second route is second-personal and can give rise to constructed, non-fact-relative reasons for action.

As explained above, second-personal reason-giving presupposes mutually acknowledged relations of authority and accountability: your claim gives me a reason for action if we each accept that you have the authority to make that claim and that I am accountable to you in this regard. In his recent review of *On What Matters*, entitled "Agreement Matters", Darwall argues that Parfit is right to seek for some sort of convergence or agreement about the normative reasons that apply. But he also argues that it is the second-personal standpoint, not the third-personal standpoint, that can account for the insight that an important dimension of normativity are "standards to which we justifiably hold ourselves and one another in common" (Darwall 2014, 104).[2]

I want to take this view on board here and continue under the assumption that some form of agreement about what we can claim from each other

2. To be clear, Darwall's claim—both in that review and in his 2006 book—is that the second-personal standpoint is necessary to account for the deontic dimension of morality. I am bracketing this stronger claim here.

can set a—second-personal—normative standard. Again, more would have to be said about what constitutes an agreement and what sort of disagreements undermine the possibility of creating second-personal reasons. This, too, will have to be done elsewhere. My modest aim here is to show how second-personal reason-giving may have independent normative force in circumstances of practical uncertainty. If it does, then Enoch's strong reducibility claim is false.

How can second-personal reason-giving help us establish what we ought to do in circumstances of normative practical uncertainty? Before I can answer this question, I need to distinguish between different scenarios, relative to our epistemic circumstances. First, it might be that we agree that there is sufficient evidence for the truth of the conditional in question. We do not know that it is true, but we can agree that I have reason to write your paper. Second, we might agree on the opposite: in light of all the evidence we have, the conditional appears false and I do not have a reason to write your paper. Again, we do not know that it is false, but we accept its falsity. Third, we might agree that the evidence is inconclusive and that the best response is that we should suspend belief. Fourth, we might disagree about what to believe about the conditional. We then need to respond to that disagreement and that response might again make suspension of belief the rational response. A fifth possible response is that we end up agreeing to disagree about the truth of the conditional.

In the last three cases, we cannot establish whether you have reason to write my paper based on our normative beliefs. But we can still establish this by agreeing on whether or not to uphold the conditional. If there is a reason for you to either write or not write the paper, it is because we can, independently of our beliefs about the truth of the conditional, agree that the conditional expresses a standard we can hold each other accountable to.

In all the cases just described, however, the answer to the question whether or not my request creates a reason for you depends on both our perspectives. Given that neither of us knows whether the conditional is true, the answer will depend, not just on our beliefs or our evidence, but also on our positive attempt to fill the normative gap created by our uncertainty, i.e. on the agreement we reach. If my request creates a reason for you, it is a constructed reason.

If we accept the possibility of constructed reasons, then we have left behind an account of reason-giving that is limited to epistemic and triggering reason-giving. The weak reducibility claim may still be true, how-

ever: some robust reason-giving reduces to triggering reason-giving. But if robust reason-giving can involve constructed reasons, then the strong reducibility claim is false: not all robust reason-giving reduces to an instance of triggering reason-giving.

This account shares with Darwall's account the idea that the second-person standpoint is linked to the construction of reasons and hence to a distinctive source of practical normativity. But as already mentioned, my account of constructed reasons differs somewhat from Darwall's original account. The difference is that my uncertainty-driven account accepts the constraint of real reasons. The place of second-personal reason-giving is residual: it occurs as a response to the normative uncertainties we face. On Darwall's account, by contrast, there is something morally fundamental about second-personal reason-giving. My account is closer to Enoch's in this regard, as I accept the possibility of fact-relative oughts, even fact-relative moral oughts. Our interactions with others can determine our normative circumstances—second-personal reason-giving matters—but so can normative reasons all by themselves.

Acknowledgements

I am most grateful to David Enoch for extensive and very helpful written comments on an earlier version of this paper and for e-mail discussions of the main claims I am trying to make. I have also benefitted greatly from Herlinde Pauer-Studer's and Kimberley Brownlee's suggestions.

References

Darwall, Stephen 2006: *The Second-person Standpoint*. Cambridge: Harvard University Press.

— 2014: "Agreement Matters." *Philosophical Review* 123(1), 79–105.

Enoch, David 2011: "Giving Practical Reasons." *Philosopher's Imprint* 11(4), 1–22.

— 2014: "Authority and Reason-Giving." *Philosophy and Phenomenological Research* 89(2), 296–332.

Kiesewetter, Benjamin 2012: "A Dilemma for Parfit's Conception of Normativity." *Analysis* 72 (3), 466–474.

Korsgaard, Christine 1996: "The Reasons We Can Share." In: Christine Korsgaard (ed.), *Creating the Kingdom of Ends*. Cambridge, MA: Cambridge University Press, 275–310.
Parfit, Derek 2011: *On What Matters: Volume 1*. Oxford: Oxford University Press.
Pauer-Studer, Herlinde 2010: "The Moral Standpoint: First or Second Personal?" *European Journal of Philosophy* 18(2), 296–310.

Grazer Philosophische Studien
90 (2014), 35–62.

MISSING THE "WE" FOR ALL THOSE "YOU'S" DEBUNKING MILGRAM'S *OBEDIENCE TO AUTHORITY*

Hans Bernhard SCHMID[1]
University of Vienna

Summary
This paper discusses Darwall's interpretation of Milgramian "obedience to authority", in which second-personal norms, second-person authority, and the power of (second-personal) address play key explanatory roles. A series of arguments against this reading is presented, and a different view is suggested, according to which second-personal authority and address have very little explanatory power. Important parts of Milgramian obedience have to be understood in the light of the human ability to look at cooperative ventures from a shared point of view. Some consequences for a more adequate understanding of the relation between the second-person and the first-person plural standpoints are explored.

Stanley Milgram's obedience experiment—in which test subjects proved to be willing to comply with an authority figure and to administer potentially deadly electroshocks to another person—is probably one of the most famous experiments in all of the history of science, and it has become part of general knowledge. It hardly needs to be summarized here (Milgram's own book-length account—after the original publication of his results in 1963—is in Milgram 1974; a detailed description of the background history and the setting is in Blass 2004, chaps 5–7). Explaining the surprising and indeed shocking degree of obedience to authority has always been recognized as an important task of moral psychology. Stephen Darwall takes up this task in a passage of the second part of his *Second Person* (Darwall 2006, 160–170). This passage seems to constitute an application of—or perhaps even a kind of "reality check" for—the account of practical reasoning Darwall has developed in the first part. The Darwallian concep-

1. I am grateful to the participants of the workshop on Darwall's *The Second Person* at the University of Vienna—especially to Stephen Darwall—and to an anonymous referee for critical comments.

tual tools—concepts such as "second-personal reasons", "second-person authority", "second-personal norms", and, above all, "address"—are used for the purpose of a description and explanation of the shockingly obedient behavior of Milgram's test people. The core critical claim of this paper is that the results of Darwall's "reality check" (if it is thus adequately described) are not nearly as favorable to his account as he seems to think. There is serious reason for doubt that Darwall's second personal approach is indeed helpful in providing an adequate description—let alone an explanation—of Milgramian obedience. Darwall's account is largely inadequate, or so it will be argued. This discussion leads to a conjecture about what might be wrong with Darwall's general account: While large parts of Darwall's analysis in the *Second-Person Standpoint* are devoted to an analysis of cooperation and cooperative-mindedness, he does not take seriously enough the fact that cooperative-mindedness involves a shared (or first person plural) standpoint from which cooperating individuals reason and act.

The paper is divided in three parts. In the first section, I shall present Darwall's interpretation of Milgramian "obedience to authority", in which second-personal norms, second-person authority, and the power of (second-personal) address play key explanatory roles. The second section will present a series of arguments against this reading, and a different view will be suggested, according to which second-personal authority and address have very little explanatory power as such, and according to which important parts of Milgramian obedience have to be understood in the light of the human ability to look at cooperative ventures from a shared point of view. The concluding section of this paper will briefly explore some consequences for a more adequate understanding of the relation between the second-person and the first-person plural standpoints.

I.

Darwall's account and interpretation of Milgram's experiments and their results emphasizes a feature of the behavior of Milgram's test people that has attracted many interpreters' attention at least since the time of Erich Fromm's *Anatomy of Human Destructiveness* (Fromm 1973, 47–52). Over the course of the experiment, the typical Milgramian test person went through intense internal and external conflicts. Of the ten test people whose cases Milgram describes in some more detail in his book, it is only one person (whom Milgram calls Bruno Batta) who complies in cold

blood. The other nine obviously struggled with themselves, and also with the "experimenter" (who gave the instruction and monitored the staged learning experiment), often engaging in prolonged and rather heated arguments. Milgram reports that his test people were shaking and sweating, voicing their worries, and demanding to break off the experiment when confronted with the "learner's" signs of distress. So "Milgramian obedience" is typically not smooth and automatic compliance, or *Kadavergehorsam*. Rather, the test people put up a rather formidable fight—which does not, however, alter the final result: the majority of the test people finished the experiment fully obedient to the experimenter's order to "punish" the "learner's" mistakes in the staged learning experiment with electroshocks of increasing intensity.

The fact that there is considerable resistance from the part of the test people makes this result all the more puzzling. After all, there does not seem to be much to overcome in terms of coercive force from the experimenter's part. He does not threaten the test people, or even scream at them, or exert some such psychological pressure. Strictly speaking, he does not even give proper orders. Rather, he first politely begs the test people to continue, and in the case of continued resistance from their part he informs them calmly that the experiment has to be continued "whether the learner likes it or not" (Milgram 1974, 16). Why would people not simply break off, at the "learner's" protests? If they really were genuinely disinclined to continue, as their repeated argument with the experimenter shows: why, then, did these "ordinary people" follow through with the experiment and perform actions which they themselves believed to be threatening to the health of the "learner"?

In his treatment of this question, Darwall heavily relies on Allan Gibbard's interpretation of Milgramian obedience as developed in his *Apt Choices, Wise Feelings* (Gibbard 1990, 58–61). Since a lot of what Darwall has to say about the variations of the experiment in which he is interested depends on this interpretation, a word has to be said about Gibbard's view before coming to Darwall's own account.

Gibbard argues that the conflict of the typical Milgramian test person, and its apparently unlikely resolution, is a case of *weakness of will.* This general line of interpretation has much going for it, and even though there is no room here to compare it to competing lines of interpretation (cf. Schmid 2011, chap. 3), some remarks on the paradigmatic Milgram test person may serve as supporting evidence. This person goes by the name Elinor Rosenblum in Milgram's book; hers is the case study to which Milgram

devotes more space than to any other of his ten case studies (Milgram 1974, 79–84). Rosenblum's case is paradigmatic because she was obedient right to the end (and did not break off, as a minority did), but—unlike Bruno Batta—she experienced intense conflicts during the experiments. Relatively early on in the experiment, when she is confronted with signs of distress from the side of the "learner", Rosenblum hesitates, reconsiders, and comes to the conclusion not to continue—Milgram reports her muttering to herself during much of the experiment, and in the debriefing, Rosenblum explains what she was saying: "I'm not going to do it. Sorrry. I'm just not going to do it." However, she does not act accordingly, and thus goes against her own judgment. Ever since Aristotle's analysis, failing to act in accordance to one's judgment has always been pointed out as the core feature of weakness of will. The view that Rosenblum's obedience may be weak-willed is further corroborated by the fact that in the debriefing after the experiment, Rosenblum explicitly states that she did not want to do it, and that she proceeded against her will (Milgram 1974, 83). This also comes with the typical difficulty of the agent to make sense of the weak-willed action as highlighted by Donald Davidson in his influential account (Davidson 2001 [1970], 43). During the experiment, Rosenblum does not understand why she is doing what she does: "for what reason am I hurting this poor man?", she keeps asking herself (Milgram 1974, 83).

Gibbard has an answer that explains why she acted as she did. According to him, Rosenblum is "in the grip" of the norm to conform to an authority's demands. The norm's grip is a sort of "powerful social motivation" (Gibbard 1990, 57f.) that bypasses the agent's own better judgment. Gibbard illustrates the grip of a norm with a series of everyday examples: "We are paralyzed by embarrassment, or a desire to ingratiate, or some other motivation that is peculiarly social. Examples abound: I may be unable to get myself to walk out of a lecture, even though it is important for me to be somewhere else. I may find myself unable to say something I know will be painful to my listener, even though I think it needs to be said." According to Gibbard, Milgramian obedience involves the same sort of paralysis: "recall the subject accepts norms against inflicting pain and danger, and accepts them as outweighing all other norms in the circumstance. These norms prevail in what he would say and think away from the scene. The norms he most strongly internalizes, though, say to do one's job, and so those are the ones that prevail in the heat of social encounter."

Gibbard contrasts the "grip" of a norm to the normal way in which norms guide our behavior, namely *accepting a norm*. If we are in the grip

of a norm, we are not rational; only *accepting* a norm belongs to the norm-expressivistic analysis of "rational" that is the core topic of Gibbard's book. Gibbard makes this point rather forcefully:

> What, after all, does a subject in one of Milgram's obedience experiments think is rational to do? If his plight is genuinely one of 'weakness of will', that is presumably *because he thinks that it makes no sense to cooperate, but finds himself cooperating nevertheless*. In other words, he does what he thinks is irrational. Now what he actually does, in this case, is a matter of the norms that have him in their grip–norms of politeness and cooperativeness. What he thinks is rational to do, on the other hand, is what is required by norms against inflicting pain and danger—and these are the norms he accepts as having most weight in his situation. In this case, then, his thinking it irrational to cooperate apparently consists of his *accepting*, as having most weight in his circumstances, norms that turn out to prohibit cooperation. Thinking something rational or irrational thus seems to be a matter of not being in the grip of norms, but of accepting them. (Gibbard 1990, 60)

Gibbard argues that the fact that Milgram's results are so shocking is due to a perspectival difference. As outside observers, we simply cannot *see* or *feel* the "grip" norms have on the participants, and we thus find it utterly surprising that the majority of the test people should cooperate. Observers tend to over-rationalize our relations to norms, thinking that it is only by means of *acceptance* that norms guide our behavior. Observers tend to ignore the power of a norm's grip, that is felt from the participants' perspective.

One may think that the swift transition from cases such as not attending to more important business because one does not want to embarrass the speaker of a lecture to which one is attending to the case of potentially kill a person because one does not want to appear non-cooperative is a bit quick. The general idea of the grip of a norm may be intuitively plausible in Gibbard's everyday examples, where the cost of not doing what one thinks is best to do is rather low. But what sense can be made of the grip of norms in cases where the most fundamental values are at stake? Why should we be prone to such irrational behavior as to go against what we think to be basic for our social life? Gibbard's interpretation concludes with an evolutionary explanation. According to Gibbard, being in the grip of a norm is another biological coordination mechanism, alongside norm acceptance, so that we have two "competing systems of control", which are both adaptive (1990, 61ff.). As applied to the case of Milgramian

obedience, however, an obvious question arises. How come Milgram's test people are in the "grip" of the norm to do what one is told to by some more or less coincidental "authority" rather than being in the "grip" of the competing norm not to hurt and kill perfectly innocent people without any good reason, in case of an error of judgment? Wouldn't it make much more sense, even from a purely evolutionary perspective, for us to be in the "grip" of the more fundamental norms that require very strong reasons to hurt another member of our community? Why should the "lesser" norm be more powerful? Gibbard does not address this issue.

Bob Solomon (2003, esp. p. 153) gives a Gibbardian-minded answer (even though he does not refer to Gibbard explicitly and casts his comments in terms of virtues and values rather than of norms). Our *acceptance* of the values of humanity, compassion, and basic decency, according to Solomon, tends to be trumped by the "grip" of obedience and conformity to authority for the following reason:

> The disposition (virtue) that is most prominent and robust in this very contrived and unusual situation, the one that virtually all of the subjects had been brought up with and practiced everyday since childhood, was doing what they were told by those in authority. Compassion, by contrast, is a virtue more often praised than practiced, except on specially designated occasions (giving to the neediest at Christmas time) or stretching the term to include such common courtesies as restraining one's criticism of an unprepared student or letting the other car go first at a four-way intersection. (Solomon 2003, 153)

Solomon's point seems to be: obedience is simply *more real* in our lives than compassion. We're just more *conditioned* to be obedient in our everyday life than to be compassionate. Thus it is the former norms or values that have us "in their grip". In this view, the likes of Elinor Rosenblum—that is, ordinary people in societies such as ours—are used to carry the values of humanity, sympathy, empathy, and kindness mostly on their lips, exercising it only on rare occasion, while their habitualized ordinary life is structured by a different set of values and norms.

> It seems to me that what the Milgram experiment shows–and what subsequent events in Vietnam made all too painfully obvious–was that despite our high moral opinions of ourselves and our conformist chorus singing about what independent individuals we all are, Americans, like Germans before them, are capable of beastly behavior in circumstances where their practiced virtues are forced to confront an unusual situation in which unpracticed efforts are required. (Solomon 2003, 153)

What Milgramian obedience reveals, according to Solomon, is a rift between the semantics of self-understanding and the structure of everyday social reality. The former values are a matter of the self-image of creatures whose real everyday life is of a totally different sort.

Looking at what we learn about our paradigmatic Milgram case, however, this harsh diagnosis seems a bit unfair. In a passage of her testimony about her everyday life quoted in Milgram's report, Elinor reports her various volunteer engagements, among them her work with high-school dropouts. She emphasizes repeatedly that in this work, as in her relation to her daughter, she believes in "love and kindness" rather than in punishment (Milgram 1974, 81). Therefore, I do not think it is adequate to suggest that in Elinor's life, values such as kindness and compassion are only a matter of lip service, as Solomon does. Similarly, there is no reason to think that the intense signs of empathetic distress which Elinor Rosenblum and her likes show during the experiment (Milgram reports that she is shivering and sweating, and that she has difficulty "keeping her composure"), her sympathy with the "learner", is just a matter of the semantics of self-description, without any connection to the hard facts and real structure of her real social life.

Darwall, in his treatment of the Milgram experiment, gives a different—and apparently more convincing—answer to the question of why it is the norm of cooperation that should hold Milgram's obedient test people in its grip rather than the norm of basic humanity. In order to see how Darwall's account works, it is necessary to look at the four variations of the Milgram experiment on which his interpretation is focused. Milgram calls the variations the "Proximity Series", and they include the following conditions:

a) Remote condition: In a previous setup of the experiment, the only way in which the "learner" made himself heard during the experiment from his place in an adjacent room was by pounding against the wall between the shock levels of 300 and 315 volt (the scale of shock levels[2] increased by 15 volts up until the level of 450 volts). The "experimenter" was placed in the same room with the "teacher" (i.e. the true test person of the experiment).

b) Voice feedback condition: In a second setup—this is the version that has become most famous and can be seen as the "standard version"—vocal

2. The electrotechnical aspects of the staged experiment are dubious. Without the corresponding watt level, no volt level gives much information about the intensity (and therefore dangerousness) of an electric shock.

protests of the "learner" could clearly be heard (presumably through the thin wall between the rooms rather than through the intercom by which the "teacher" posed the "learning experiment"-questions; the protests were played from an audiotape). Darwall gives only a rough description of this version, but it shall turn out to be useful to give some more details here. In this version, the "learner" voiced first signs of pain at 75 volts, he started demanding to be let out at 120 volts (unfortunately, Milgram's book does not contain the exact content of the "voice feedback" in the original condition, but he does so in his report of the "new baseline" condition, in which the "victim" repeated this demand—or order—no less than 37 times in the following rounds; Milgram 1974, 56f.). At around 300 volts, he voiced agonized screams, he declared not to be part of the experiment anymore, and that he refused to give answers. After 330 volts, he remained silent and did not answer the learning questions.

c) Proximity condition: In the third setup, the "learner" was placed in the same room with the "teacher" and the experimenter, at only a few feet distance from the "teacher", and he could not only be heard, but also be seen by the "teacher". Thus in this setting, the questions of the staged learning experiment did not have to be asked via intercom, as was the case in the previous settings (also, it is likely to assume that in order to deceive the test person, the protests had to be voiced by the "learner" live instead of being played from a tape recorder, unbeknownst to the "teacher").

d) Touch proximity condition: In the fourth and final setup, the distance between "learner" and "teacher" was reduced further, and the hand of the "learner" was not strapped tightly to the electrode so that the learner could remove his hand from the contact, which he did after 150 volt. The experiment could only be continued if the "teacher" conformed to the experimenter's demand and pressed the "learner's" hand on the electrode.

A table on page 36 of Milgram's book shows the drastic and linear drop of the obedience rate from setting a) to setting d): The mean maximum shock delivered in the first condition was 405 volt, dropping to 360 volt in the voice feedback condition, to 315 volt in the proximity condition, and finally to 255 volt in the touch proximity condition. Darwall gives this the following interpretation. The situation in which the "teacher" is placed is a conflict between second-personal norms (second-personal norms, according to Darwall, are norms "that concern the authority to claim and demand" [Darwall 2006, 153]; they differ from non-second-personal norms that concern reasons that are independent of anybody's claim or demand). On

the one hand, there is the norm of obedience and cooperation (the experimenter's authority to demand), on the other hand, there is the norm not to hurt an innocent human being (the "learner's" authority to demand). According to Darwall, the reason why the norm of obedience rather than that other norm—the basic norm of humanity—has a "grip" on the test people is that in the first conditions people are more susceptible to the experimenter's authority to demand, because as opposed to "the victim", he is close to them. Here, an important feature of Darwall's account of practical reason comes into play: "address". "Address" is a word Darwall never seems to define, but which he often uses to refer to the way a demand or claim is made to another person (e.g., Darwall 2006, 4). The upshot of Darwall's discussion of Milgram's "proximity series" is that susceptibility to address correlates to (some aspects of) physical distance or perceptibility. The idea seems to be that in this case at least, the psychological force of second-personal norms depends on the actual address of the relevant authority. This explains why the closer "the victim" is brought to the test person, the more dramatic the obedience rate drops.

More precisely, Darwall emphasizes the following features as relevant for the explanation: In the variations b (voice feedback), c (proximity), and d (touch proximity), "the victim" explicitly addresses the teacher.[3] This address, according to Darwall, plays a crucial role. In variation c, a situation of mutual awareness (common perceptual knowledge) is created. "The most significant change in the overall rate of disobedience came when the learner was moved into the same room with the subject. The subject could then see the effects of what he was doing. But also importantly, he was aware, for the first time, of the learner's awareness of him" (Darwall 2006, 166). "Bringing the learner into the room gave him a presence as someone to whom the subject was accountable—a second-personal advantage the experimenter had in all the setups. […] The other was present now not just as someone with the standing to demand compliance with a norm requiring that he not be harmed but also as someone with some standing to judge one's compliance with it" (ibid.). With proximity increases the degree of empathy. Darwall emphasizes the role of empathy (which is conceived of by Darwall along roughly Batsonian lines). Empathy is the condition in which we are likely to look at things from the other's point

3. "Clearly, having protests addressed to them made a significant difference in the subject's behavior" (Darwall 2006, 165). This quote, as well as Darwall's line of argument, shows that Darwall assumes that in the voice feedback condition, the "learner's" protests are addressed to the "teacher" rather than to the experimenter.

of view: "Empathy took them inside their victim's perspective", Darwall says about the test subjects in the proximity and touch proximity conditions (Darwall 2006, 167).

II.

What follows is a critique of Darwall's account of Milgram's obedience experiment (a) and an alternative view (b). The critical claim is that Darwall's explanation is based on at least one factual mistake and a series of empirical assumptions that are highly problematic. The alternative view is that at least for the explanation of the first stages of Milgramian obedience, we do not need to appeal to any mysterious weakness of will or power of address, but rather to a solid conception of what it means for a cooperative-minded agent to engage in a joint action.

a) *Critical Remarks.* It is not surprising that Darwall places the role of "address" at the core of his interpretation. This is the factor that Darwall quotes in explanation of the first obedience level drop in the variation series. And this is where the trouble with Darwall's interpretation starts.

1) Contrary to what Darwall claims, "the victim" does not address his orders to be let out to the teacher. In his description of the proximity variations, Milgram explicitly states that at 120 volts, "the victim shouted to the experimenter that the shocks were becoming painful", and that at 150 volts, "the victim cried out, 'Experimenter, get me out of here'" (Milgram 1974, 23). There is no indication that the "victim" changed the addressee of his protest at any point of the process.

The same is true for the "new baseline" experiments, where Milgram gives the exact wording of the voice feedback (Milgram 1974, 55ff.) In most of the 37 repetitions of the order "Let me out!", as well as in phrases such as "You have no right to keep me here!", "the learner" does not specify the addressee. "You" may either address the teacher, or the experimenter, or both. However, the learner makes clear twice in the process that he addresses his protests to *the experimenter* (who is sitting behind the teacher), rather than to the teacher. The first time he makes the target of his "address" clear is in his first demand to be let out (in the previous rounds, he had shown non-verbal signs of distress and made remarks such as "Ouch, this really hurts!"). This is the first in the long series of demands to be let out,

and it is addressed to the experimenter. Here is what the "learner" says after receiving the 150-volt shock: "Ugh! Experimenter! That's all. Get me out of here. I told you I had heart trouble. My heart's starting to bother me now. Get me out of here, please. My heart's starting to bother me. I refuse to go on. Let me out" (Milgram 1974, 56; the "victim" did not mention having heart troubles in the proximity variation experiments). Likewise, the first in the series of claims not to be part of the experiment anymore is explicitly addressed to the experimenter at 210 volts: "Ugh! Experimenter! Get me out of here. I've had enough. I won't be in the experiment any more."

How does this bear on the interpretation of the "address"? It seems to be a basic convention of conversation that competent speakers specify the target of the address when it is changed, and when the circumstances do not make the change of target obvious. From "the teacher's" point of view it is reasonable to assume that "the learner" had made it explicit, had he intended to address "the teacher" rather than the experimenter at some point of his protest. Thus it seems clear that Milgram's test people had to understand *all* of the "voice feedback" as addressed to *the experimenter,* not to themselves.

Thus Darwall is simply mistaken in his interpretation of the target of "address" involved here. Milgram's test people clearly had to understand *all* of the "voice feedback" as addressed to *the experimenter,* not to themselves. Especially if we agree with Darwall that address matters in human cooperation, this seems to be a rather fatal flaw of Darwall's interpretation. The explanation of the behavior of "the teacher" cannot be a question of susceptibility of second-personal address, because there simply is no address to "the teacher" from the side of the "victim".

It is an open empirical question—and a very interesting one indeed—how the test subjects would have acted, had the learner's address been of the Darwallian kind. What if the learner had tried to talk to the teacher instead of the experimenter? It seems plausible to assume that the effect would have been significant. In particular, it seems likely to assume that test people would have tried to respond directly to the learner first, rather than to turn to the experimenter, and to engage in a futile discussion with him, as they did in Milgram's version. Since "the victim's" address is to the experimenter, it is no wonder that Milgram's test people discuss the question of whether or not to continue with the experimenter, rather than trying to address the "victim", which otherwise would be the much more obvious move. But whether or not this conjecture is true, it is certainly a

mistake to assume that the address is to the teacher in the voice feedback condition.

2) As mentioned above, Darwall suggests that only in conditions c) and d), a situation of mutual awareness and common knowledge is created. He assumes that there is no mutual awareness or common knowledge in conditions a) and b). Whatever the conception of mutual awareness or common knowledge Darwall has in mind may be, it seems clear that for those attitudes to be mutual or common, reciprocal higher-order attitudes have to be in place or at least to be available to the participants. It is far from obvious, though, that this is indeed not the case in conditions a) and b). We simply don't know what exactly Milgram's test people were likely to assume about the "learner's" awareness or knowledge of their own attitudes and actions. Darwall seems to think that no such reciprocal awareness could be assumed by the test people in a) and b), because in these experimental setups, the teacher and the learner could not *see* each other, since they were placed in different rooms. Tacitly assuming that mutual awareness needs to be of the visual kind, Darwall concludes that no mutual awareness could have been assumed. However, voice allows for a sort of mutual awareness, too, and the question of acoustic mutual awareness seems open at least in condition b). The decisive empirical question here is: did the test person assume that the "voice feedback" condition was one of mutual acoustic awareness? E.g., did he or she assume that the victim could hear what he or she was saying to the experimenter on the victim's behalf when they were arguing with him and demanding that the experiment be terminated? As we shall see below, this question matters a great deal for the interpretation of the ensuing behavior, and there does not seem to be any conclusive evidence. The video footage of his experiments which Milgram released, and which seems to focus on the voice feedback condition on the "new baseline", has "the victim's" voice come over very clearly from the adjacent room. If that's how the teacher heard the learner's voice, we would have to assume that the teacher took the situation in condition b) to be one of acoustic common awareness, assuming that "the victim" could hear him or her just as well as he or she could hear "the victim" (one test person in the movie tries to address the learner verbally, asking if he's o.k., obviously assuming that there was acoustic mutual awareness of the kind that allows for verbal communication). Like the first critical point above, this hinges on an empirical question, and in this second case, I grant that it is an open one. If the answer should be in the positive—and there is at least *some* evidence that it might be, given the passages of Milgram's

movie just mentioned—, common awareness cannot be appealed to in explanation of the obedience rate drop between conditions b) and c), as Darwall does. A related point is that Darwall claims that in condition b), the teacher assumes the "victim" not to be in a position to judge over whether or not his complaints have an effect; even if we grant that there is no mutual awareness in condition b), it is hard to see why this should be the case.

3) Darwall appeals to empathy as an explanatorily relevant factor in his interpretation of the drop of obedience between b), c), and d) (as is obvious from the above quote). Empathy, he seems to claim, comes into play in condition c), and it is reinforced by increased proximity in condition d). This line of argument seems to imply that empathy plays a lesser role in the previous versions, and that the test people are somewhat less concerned with the "victim's" apparent suffering in the voice feedback condition. Given what we learn about the likes of Elinor Rosenblum who is acting in the "new baseline" version of the voice feedback condition, however, it does not seem plausible to assume that there is a lack of empathy or a less-than-full awareness of the "victim's suffering" at all. Milgram reports that her considerable stress is only a sign of her worries for herself, and not an expression of any empathetic concerns, but that seems a bit arbitrary, to say the least. Elinor Rosenblum and her likes obviously do not need to *see* their "victim" to empathize with him; hearing his "voice feedback", and knowing about the effects their actions have on him, obviously suffices, and they are fully aware of "the victim's suffering".

4) Last, but not least: As mentioned above, Darwall sees the Milgramian test person as torn between two second-personal norms. It seems hard, however, to ignore that the respective claims made on the behavior of the teacher by the learner and by the experimenter, respectively, are of a very different kind. The "victim" voices the demand not to be hurt in his physical integrity, while the experimenter's demand is to continue with a "learning experiment" that is aimed at finding out about the influence of punishment on learning performance. Darwall seems to reduce the question of Milgramian obedience to a kind of vector geometry of opposing addressive pushes or pulls working on the psyche of the test person: the closer the "victim" is placed to the "teacher", the more effective his address becomes against the countervailing power of the experimenter's address. But the demand not to hurt a person and the demand not to cause any trouble for an ongoing procedure cannot be on a par, even on a purely psychological level. "Don't kill me!" and "Do your job/Don't

make my job difficult!" certainly cannot be assumed to have the same psychological power. Even when the proximity is roughly equal, one third of the test people still give more weight to the experimenter's demand—why? It seems that in Darwall's interpretation, this central question remains open.

Before coming to a different interpretation, some remarks concerning the epistemology of Milgram's obedience experiments are in order. Recall Gibbard's claim (compatible with Darwall's view) that what makes Milgram's results so unexpected (and indeed shocking) may simply be a matter of the perspectival difference between observers and participants. As observers, we like to believe that we would never ever go along with the experiment, were we in the test people's shoes. In his own presentation of his results, Milgram reports that he had done the same with professional psychologists before the actual experiments himself, with the result that none of the experts expected that the experiment would ever work out, as they assumed that rarely any test person would go along (Milgram 1974, 27ff.). Given the experimental results, however, we have to learn the (apparently hard) lesson that it is very likely that in actual fact, we would have acted in exactly the same way as the large majority of Milgram's test subjects. Gibbard's conclusion is that there is something about Milgram's experiment which makes observers unable to *see,* as observers, what motivates the participants.

There is empirical evidence that supports the claim about the importance of the perspectival difference. At a time at which Milgram's experiment had not yet attained its status as a part of general knowledge, the psychologist Günter Bierbrauer (1979) carried out an experiment in which the test people were informed about the setup of Milgram's obedience experiments and then had to estimate the level of obedience. Bierbrauer presented the setup of Milgram's experiments to his test subjects in the same way Milgram himself did with his fellow psychologists before he started the actual experiment, only that in Bierbrauer's case, a second group of test subjects had to think, discuss, and to write about the experiment before coming up with an estimate of the expected compliance. Lo and behold, this second group came up with a remarkably more realistic estimate of the expected level of compliance than the group of test people who just had received the information about the setup of the experiment.

What thinking about the experiment and discussing it with others enables people to do, is to cross the perspectival hiatus between observer

and participants. A decisive question therefore is: what exactly is it that those people can see that we, as mere observers, cannot?

Gibbard thinks that the decisive factor here is *the grip of the norm of cooperation,* and Darwall refines this line of interpretation with his view that it is really the power of the experimenter's *address* that plays this role. The following is a different answer. What we can't see, as observers, is no mysterious "grip of a norm" or "power of address"; it is very simply that the participants look at the experiments, and reason in the experiments, *as participants,* that is, from the point of view of a person engaged in what he or she takes to be a joint action, together with the "learner" and the experimenter, and that from this viewpoint, there is *reason* to continue. The viewpoint of a cooperative-minded person is first-personal, in the plural, and it is this feature that we tend to miss when we look at the matter as observers. As opposed to Darwall's emphasis on the second person and on Gibbardian weakness of will, the key to bridging the perspectival gap is in the first person plural standpoint and in the reasons there are for cooperative-minded agents.

b) *An Alternative Account.* The idea to interpret Milgram's result as a consequence of the nature and structure of joint action is not entirely new. Margaret Gilbert, whose account of the first-person plural perspective will be crucial for the following, has an unpublished manuscript on Milgram's experiment which highlights the (assumed) *joint commitment* between the participants (Gilbert 2010). From the perspective of Milgram's test people, the test person has, together with the "learner" and the experimenter, entered into an agreement to perform a learning experiment. The central claim is that to look at a venture as a joint action has consequences for what a person feels he or she *ought* to do.[4] A person who does not understand that being part of a joint commitment has normative consequences is not a cooperative-minded person. To use Gilbert's favorite example: you don't understand what it means to engage in a joint action if you think you can simply walk away from it whenever you decide or believe

4. An anonymous referee suggested that the following analysis misinterprets the Milgram experiment as a joint action of the wrong type, and that the cooperation in place should be understood as weak I-mode cooperation rather than strong we-mode cooperation. I do not use or discuss Raimo Tuomela's distinction in this analysis. Contrary to what the referee seems to be thinking, however, the following analysis of cooperative-mindedness does not take the participants to form a group, independently of the action, but rather something like a Gilbertian plural subject of the action, that is, as a group only in the sense of those acting together.

to have an independent reason to do so, without at least owing the others an explanation. This is a first and basic feature of what it means to be cooperative-minded:

1. A cooperative-minded person understands that a joint commitment places him or her under some normative constraints. There is a sense in which if the action is jointly intended, he or she should do his or her part.

This is the feature on which Gilbert focuses in her account of the Milgram experiment; let us see how far this explains Milgram's result in a way that bridges the perspectival gap. It certainly explains why Milgram's test people—who obviously do not like what is going on, and who would rather stop the experiment—do not simply walk away. However, there are two reasons to doubt that this takes us very far. In her manuscript, as elsewhere in her work, Gilbert fully acknowledges that normativity is not exhausted by the obligations created by joint commitments. It seems that even if the participants accept that there is some normative pressure towards "cooperation", there seem to be very clear, overriding moral reasons against continuing the experiment. No *reasonable* person, however cooperative-minded he or she may be, assumes that the normative reason created by the joint commitment to carry out a learning experiment justifies electrocuting a person; any such reasoning would simply be insane. So there appears to be a tension between the first principle and cooperative-mindedness and a basic restriction placed on cooperation by moral reason:

2. A *reasonable* cooperative-minded person understands that if there are reasons to assume that the joint action is bad, the joint action should not be carried out. If the joint action should not be carried out, he or she should not do his or her part.

Thus the basic question remains: granted that they are cooperative-minded, why would Milgram's test people be so *unreasonable* not to be sensitive to this restriction—especially since Milgram's report makes it clear that the test people were clearly aware of the moral problem? Why was their cooperative-mindedness so *unreasonable*? It seems that the gap remains. How can it be closed?

Cooperative-mindedness has another feature that becomes important here. Thinking about the basic question suggests a certain view of nor-

mative conflicts that needs to be called into question. We usually tend to think of moral questions as internal affairs, somehow to be resolved within the individual's own mind (e.g., the internal voice of conscience against equally internal desires, or some such). Yet this is clearly not how normative conflicts are resolved by cooperative-minded persons. Rather, the following tenet of *reasonable* cooperative-mindedness comes into play:

> 3. A reasonable cooperative-minded person understands that the reasoning about whether or not the joint action is good or bad should be carried out communicatively. If a cooperative-minded participant comes to the conclusion that the joint action is bad, he or she should not just walk away, but voice her concerns, discuss them with his or her partners, and aim at reaching a joint decision to stop.

Provided that it is not obvious to the test people that the alleged "experiment" is bad (after all, it is presented to them as an important learning experiment), it is very much in line with the demands of reasonable cooperative-mindedness that Milgram's test people should not just stop and leave at the point at which the "learner" shows signs of distress, but rather interrupt the "experiment", tell the other participants that he or she thinks the joint action should be aborted, and discuss the issue with them in order to reach a joint decision to stop. In Gilbert's account, this corresponds to the view that the dissolution of a joint commitment can only be performed jointly. The decision to stop should be a joint decision, based on consensus. In joint action, consensus plays the same role as individual resolve does in the case of individual action. That's exactly what Milgram's test people are trying to achieve in entering into a discussion with the experimenter and in voicing their concerns and views. However, instead of engaging in proper joint moral reasoning, the experimenter simply responds with his "prods".

Here, the explanatory power of reasonable cooperative-mindedness seems to end. It is obvious that the fact that the partners are not convinced by one's concerns and think the joint action should be continued does not license a reasonable cooperative-minded participant to continue—at least if the others do not present convincing arguments for their view, showing that the joint action is not bad after all. In Milgram's case, the experimenter says to the protesting test person that he thinks the experiment should be continued, but he does not present *any* argument for his view (with the exception of the somewhat dubious "special prods" that "the experiment

requires that you continue", and that "although the shocks may be painful, there is no permanent tissue damage"; Milgram 1974, 21). Thus it seems to be in blatant violation of a further feature of reasonable cooperative-mindedness that the majority of the test people continue:

> 4. A reasonable cooperative-minded person understands that if no rational consensus can be reached on whether or not the joint action should be continued, and if he or she is convinced that the joint action is bad, he or she should desist from doing her part.

Herbert Clarke (2006) is another author who has approached Milgram's obedience experiment from a distinctly joint action-centered point of view. One important element of his interpretation is in the analysis of the structure of accumulating joint commitments over the course of the experiment, a feature that is very much in line with Gilbert's view. However, there is another important factor that Clarke highlights. Clarke emphasizes that from the perspective of cooperative-minded test people, the "prods" with which the experimenter "replied" to their concerns in the discussion—such as "please continue", "please go on", "the experiment requires that you continue", uttered in a calm and almost bored tone of voice—must have been interpreted by the test people as conveying an implicit message, perhaps along the lines of: "there is no reason for your concerns", "your view that the learner is in serious danger is so absurd that I won't even dignify them with a proper reply". Thinking further along these lines, an important factor comes into view. Cooperative-mindedness does not only commit to joint reasoning in normative matters. Not just the question of whether an action is good or bad should be answered in joint reasoning, where this is possible. The same applies to *epistemic* questions. What the matters of fact are, and what should be taken to be the case should be settled cooperatively, too.

> 5. A reasonable cooperative-minded person understands that questions of matters of fact should be answered cooperatively, where this is possible.

This means that a reasonable cooperative-minded person understands that other people may be competent cognizers, too, and perhaps more competent than him- or herself, depending on the circumstances. Independently of which theory of the epistemic role of discourse and testimony one

chooses to adopt, it seems clear that according to any sound conception, a reasonable cooperative-minded person will not always stick to her own epistemic interpretation of a situation, no matter what other people seem to think. He or she has a reasonable view of who's likely to be in the know, and he or she is ready to adapt his or her own interpretation accordingly. How does this bear to the case of Milgram's test people? It does not seem unlikely to assume that they might have understood the experimenter's prods in the early stage of the experiment (with the learner reacting with "ugh", "ouch, that really hurts") as conveying an implicit message somewhat along the lines of: "Don't worry, this is not what it seems to you, we're quite used to the fact that the learner 'loses it' for a moment when he receives a shock." We have to consider the possibility that the test people may have taken the experimenter's prods into account as evidence concerning the question of what it really is that is happening: is the learner really in danger, or does he just "lose it" briefly upon receiving the shock? It is true that in the debriefing after the experiment, Milgram's test people reported that they believed the learner to be in serious danger, an issue that needs a more thorough discussion (cf. Schmid 2011, chap. 6).

Let us focus on another problem here. Whether they believed the "learner" to be in serious danger, or whether they took the experimenter's statements as saying that the matter may actually be not as serious as that, it seems clear that no reasonable cooperative-minded person would cross the line drawn by the learner when he demands to be let out.

> 6. A reasonable cooperative-minded person understands that one party's demand to abort the joint action is *in itself* a very strong reason to agree to end the joint action.

This is especially obvious in the kind of joint action at stake in the staged learning experiment, where the learner has agreed to be strapped to an electric chair and now demands to be let out. The reason why such a demand is a reason in itself, i.e. independently of the reason the learner may have for his demand ("this hurts!"), is in the (thinly) voluntary nature of joint actions. Joint actions are joint actions only if the participants want to perform them (whatever individual or shared reasons they might have for wanting to perform them), that is, only if the participants participate *knowingly* and *intentionally*. It seems clear that in the given case, the joint action has simply lost its basic consensual base, and the "victim" in his electric chair is being *made to* "participate" rather than actively participating, and

the depressing fact of the matter seems to be that Milgram's test people are not reasonable cooperative-minded people, since they do not stop playing their role in the joint action; the shocking gap seems to remain.

However, there is one reason for doubt. Take the voice feedback condition, and imagine the situation from the point of view of, say, Elinor Rosenblum. You have agreed to be part of a learning experiment and to play the role of the teacher. After a few rounds, the learner shows signs of distress: "Ugh, that really hurts!", you hear the learner saying. Perhaps you try to talk to him through the wall, or via your microphone (as one test person can be seen doing in Milgram's movie): "Are you all right in there?" You don't get an answer. So perhaps you turn to the experimenter, and you voice your worries. You may say something like "I think we should stop and see if he's all right." The experimenter simply replies "Please go on!", in a very calm voice. You take this to mean: "There is absolutely no reason to worry. Test people very often show brief signs of distress, but that does not mean that they have a serious problem. Your concerns are so absurd that I won't even start a discussion." You don't just believe him, but you are now uncertain to some degree. This is certainly not the way it looks to you, there were clear signs of distress. And there are these worrying labels on the board of the "generator" with which you deliver the shocks, where it says "severe shock". But after all, there is the possibility that the experimenter might be right. He implicitly tells you that there is no real distress. And why would they be putting the participants' life in danger for the sake of a learning experiment? In addition to that, another thing slowly starts to register in the back of your mind: during your conversation with the experimenter, you haven't heard anything from the learner. Assume that you take this to be a situation of acoustic common knowledge. Why does he not say whether or not he is ready to go on? This becomes significant. So your question is: is the learner ready to continue or not? As you don't get any answer from him, the only way to find out is to ask the next learning question. If he's unwilling to continue, he will surely not just give an answer (and thereby risk an even stronger shock). If he does, he seems to be willing to go on, thereby corroborating what you take to be the experimenter's interpretation of the situation. This seems to be a reasonable way to find out whether or not it is o.k. to continue with the experiment, and in fact the only way left open to the test people by the rather Machiavellian setting of the "real" experiment. Lo and behold: the learner does answer the question. So you think he's o.k. with continuing the experiment after all. After the next punishment, he voices signs of

increasing distress, but the experimenter's reassuring remarks as well as the learner's continued smooth cooperation are again reasons to take this not to mean that the experiment should be terminated. So you go on, in spite of the learner's "ughs" and "ouch, that really hurts".

After a few more rounds, however, the situation changes. The learner now addresses the experimenter, and explicitly demands of him to be let out. Naturally, you now turn to the experimenter—surely, he will declare the experiment to be terminated now! Much to your surprise, however, he's not showing any reaction at all. You find this rather weird; how can he not bother about that very legitimate claim that clearly shows that the experiment is over now? Has he not heard what the learner just said? So you repeat the learner's point to the experimenter. He calmly informs you that the experiment continues. You take this to mean something like the following: "It is perfectly normal for test people to 'lose it' for a moment when they receive an electroshock. Contrary to what they may claim, that does not mean that they're not ready to continue, let alone that they're in serious danger." You're not so easily convinced, you insist, the discussion with the experimenter becomes more heated now, at least what your part in it is concerned. He, however, keeps his calm and bored attitude. And again, you start to register that you haven't heard from the learner since his reaction to the shock. It is likely that you assume that he can hear what you're saying to the experimenter. His silence is very strange. Why would he not repeat his protest and support you in the advocatory role you're playing on his behalf? Why does he not say something like: "Experimenter, the teacher is right; we have to stop here!" Might it be the case that in spite of all the evidence to the contrary—the learner's screams and protests, as well as the alarming labels on the volt scale—the experimenter is right? Has he just lost it for a moment, or is he in serious trouble? You're undecided. Is the learner ready to go on after all? How can you find out? Perhaps you try to address the learner directly, but again, he would not answer. Again, the only thing you can do to find out is simply to ask the next question, assuming that if he's not willing to continue, as he just said, he will certainly not answer your question. But much to your amazement, he does, thereby indicating that he's willing to go on after all, very much against what he just stated.

A third dramatic change happens some rounds later: the learner now explicitly states that he's not part of the experiment anymore, and that he refuses to give an answer. And in the last stage of the experiment, there's nothing to be heard from him—he may have fallen unconscious.

I have no explanation for "collaborative" behavior at this last stage. But I like to think the account I have given points at important explanatory factors for the previous stages, and whatever the explanation for the last stages may be, it seems that this may matter (for a detailed version of the argument cf. Schmid 2011). As to the stage of the experiment up to 330 volts, it seems that the explanation for the behavior of the obedient test subjects need not appeal to weakness of will, lack of empathy, second-personal address, or any such factors. What is needed, however, is an adequate conception of cooperative-mindedness: of the kind of practical reasoning in cooperation. In the account I have just given, the test person acts on the basis of an understanding of the experiment as a joint action with the shared goal of finding out more about the nature of human learning. The test person assumes that there is a mutual agreement to carry out that experiment, based on each participant's willingness to take part. That mutual agreement creates a joint commitment that has some normative weight: if there is no stronger reason to discontinue, each *should* perform his or her role. Just leaving without any previous discussion is not an option for a cooperatively competent person. However, it is clear that there may well be very good reasons to discontinue, and clearly, a participant's being hurt is a strong case in point. Faced with the learner's signs of distress, the task ahead for the test person is a twofold one: What is happening? And what should be done? The first question is how these signs should be interpreted: are they a case of "losing it" for a second, or do they mean that the learner wants to discontinue the experiment? As a competent cooperator, the test person will only continue if the learner is ready to continue. And built into the experiment—yet not into the usual interpretation thereof—are signs which the test person will interpret as expressions of the learner's readiness to continue.

The way in which the Milgram experiment is usually depicted suggests that it must be obvious to the test person that a) "the learner's" position is clear and that b) the experimenter is unwilling to respond sensibly to objections, so that c) a reasonably cooperative-minded test person should walk away. The reconstruction given above, however, shows that this is not the case. Up to the level of 330 volts, compliance is not in violation of cooperative-minded practical reasoning. Thus the reason for the test person's compliance is not to be seen in some weakness of will, or in the power of the experimenter's address, as Gibbard and Darwall have it, but rather in cooperative-minded practical reasoning, that is, an evalua-

tion of the situation based on an understanding of the normative reason constituted by a joint commitment, together with the assumption that the experimenter is a trustworthy epistemic authority concerning the interpretation of the situation whose view should be given some weight in the evaluation of the situation, and an understanding that the experiment should be discontinued if the learner's will is firmly set against it.

The main difference between the kind of practical reasoning appealed to in this interpretation and the Darwallian version is that "address" does not play an independent role; it is not the susceptibility to second-person authority that explains the behavior, but rather cooperative reasoning along the lines sketched above. In order to understand the practical weight which the test person assigns to the statements uttered by "the victim" and the experimenter, respectively, it is important to see that the test person reasons from a we-perspective, as it were.

Comparing the second-personal approach, as developed by Darwall, and the first person plural view as sketched above, the decisive question is how each fares in the interpretation of the proximity variations of Milgram's experiments, on which Darwall's analysis is focused. How is the dramatic decline of the obedience level over the course of the proximity variations to be explained? Darwall's view is: With increasing proximity, address becomes possible and more vivid, especially since empathy, or higher degrees thereof, come into play. I think that the first person plural account I have sketched above, and the critical light it sheds on the way the Milgram experiment is usually depicted, suggests a different reading. Imagine the proximity condition from the test subject's perspective. Again, assume that what you're trying to find out is: Do the learner's protests signal a firm will that the experiment be discontinued (as his reaction immediately after the shock suggests), or rather just a momentary loss of control (as the experimenter's statements, and the continued cooperation of the learner in the next round suggests)? In this condition, the fact that the learner does not support you in your conversation with the experimenter on his behalf becomes particularly puzzling. Why does he scream to be let out when he receives the shock, but then keep completely silent when you argue with the experimenter? Even more importantly, we have no report about how the learner reacted in the proximity condition when he was directly addressed by the test subject—it seems hard to imagine that this did not happen regularly. In Milgram's movie, which is focused on the voice feedback condition, a test person can be seen trying to address the

learner through the wall—with no reply from the learner's part, so that the test subject then turns to the experimenter and voices his worries. Was the learner's "reaction" to address the same in the proximity condition, in which direct address must have occurred regularly on the first signs of distress? What sense would an average test person make of this incomprehensible oddity? I suspect that the very oddity of the learner's silence may have been an important part of the reason why compliance dropped so dramatically; people just felt in their guts that there was something deeply wrong with the whole situation (even though they reported in the debriefing that they believed that the learning experiment was real). The basic problem is certainly the same in the remote/voice feedback and the proximity/touch-proximity conditions: There is clear evidence that the learner is in distress and wants the experiment to be discontinued. At the same time, up to the level of 300 volts, there is some evidence that even though the experiment is painful to the learner, he is still willing to continue. The decisive difference, however, is that in the proximity and touch proximity conditions, there is more reason to assume that consent should be given verbally by the test person. In these conditions, asking the next question is not a particularly obvious way to find out whether the learner still wants to continue or not anymore. You would expect the learner to say whether he wants to continue or not if he's in the same room with you, so you're less likely to resort to other ways of finding out whether or not he's still "in the game".

Against Darwall's account of the behavior of Milgram's test people as weak-willed targets of address and subjects to second personal authority, it seems that in the light of the account sketched above, they should rather be seen as cooperative-minded we-reasoners. And the gap between the participant perspective and the observer perspective is that the way the experiment is depicted (and thus observed) is, to some degree, individualistic: the single actions are mentioned, without showing the way in which the test people must have assumed they formed a whole joint action. Important explanatory factors are in the *interrelations* between these actions (e.g., the test person's protest, the experimenter's "replies", and the "learner's" silence).

I should emphasize once again, however, that the reconstruction I have given works only up to the level of 300 volt, and that I have not offered any explanation for the behavior of those who went beyond that point; yet it seems that the difference in the interpretation of the initial stages does matter, even if the explanation is incomplete.

III.

In conclusion, some remarks on the conceptual bearings of the above discussion. Darwall's book on the Second-Person Standpoint is largely on cooperation. Why then, did he not approach the Milgram experiment from the point of view of cooperative-mindedness? I suspect that Darwall's difficulties with Milgram's obedience experiments are symptoms of a more fundamental point: a problem with his conception of the relation between cooperation and second-personal phenomena. More precisely, I shall argue that it is because he has no sufficiently developed account of human cooperation—and because he underestimates the degree to which the phenomena which he claims to be second-personal (second-personal reasons, second-personal norms, second-personal authority etc.) are really a matter of the first person plural—that he misconceives of the results of Milgram's obedience in the way he does. It is because he has no adequate conception of joint action and we-attitudes that Darwall depicts the Milgram experiment as a compound of individual actions and I-thou-relations, which makes it impossible to see the important explanatory factors I have highlighted. Darwall's account, as many others, misses the forest for the trees, or the alphabet for the letters; it is focused too much on the single actions and relations involved in the experiment and neglects that they play their role only as parts of the whole. This corresponds to Darwall's general tendency to depict phenomena of human cooperation as a matter of "you's", whereas it seems to me that any "you" is part of an I-thou-relation that involves some sense of "us". In this concluding part of the paper, I shall try to explore the difference at stake here.

Darwall is certainly aware of the fact that the first person plural matters. Indeed, he sometimes goes as far as to place it at the heart of his analysis. "I claim that to understand moral obligations as related to moral responsibility in the way we normally do, we have to see it as involving demands that are in force from the moral point of view, that is, from the (first-person plural) perspective of the moral community", he says in the introduction of his book (9). This reminds heavily of Wilfrid Sellar's original motive for introducing the term "we-intention". We-intentions explain how moral claims can both be expressive of *attitudes* and have a claim to intersubjective validity, too; they are, Sellars claims, expressive of *shared* attitudes—i.e. attitudes of the widest, least parochial community, but first person plural attitudes nevertheless (cf. Schmid/Schweikard 2013). However, there is

a reason why Darwall puts *The Second Person* rather than *The First Person Plural* in the title of his book. He does not see the second-personal phenomena he is describing (address, second-person authority, etc.) simply as features of the way "we" are related to each other as members of some "us". Rather, he argues, in the passage following the above quote, that the second person stance is "a version of the first person viewpoint (whether singular or plural)". Accordingly, he does not engage with the literature on we-intention, joint action, and collective intentionality—which may well have led him to a more adequate description of Milgramian obedience and the kind of reasoning that explains Milgram's results. The one exception is a brief discussion of Margaret Gilbert's Plural Subject account, which, according to Darwall, "helps clarify the relation between the second person and the first person plural standpoint" (Darwall 2006, 182). The upshot of this discussion is best captured in Darwall's slogan: "The way to 'we' runs through 'you' and 'I'" (Darwall 2006, 178). With this, Darwall concludes a first (and very short) discussion of the idea that cooperation involves joint intentions, i.e. intentions that are "ours" and that involve an understanding of the participants as some form of unity, insisting on the relevance of reciprocal recognition between the agents, and the structure of authority and address that Darwall labels second-personal.

I see at least three possible readings of that Darwallian slogan—a weaker and innocent one and two stronger, more problematic versions. In the immediate context of that claim, the following weaker reading is suggested: In order for agents to share a full-blown, agreed-upon collective point of view, they need to recognize each other in a way that Darwall calls second-personal. It is not the case that the "we", in that sense, comes *before* the kind of interpersonal relations Darwall has in mind. The first person plural is not the foundation of the second-personal standpoint; no "we" can be made sense of without understanding the relation between you and I.

Darwall makes this point clear in his discussion of Margaret Gilbert's Plural Subject Theory. Gilbert (1989 ff.) has gone to great lengths arguing that it is important for the understanding of the social world and the structure of mutual obligations to recognize the importance of the "first-person plural" perspective. In this view, it is the *unity* between people that is at the core of human cooperation, the "pooling of wills", and the joint intending "as one body". Yet the formation of a plural subject involves interaction between the participants—some sort of agreement has to be reached, and this procedure involves Darwallian themes. Darwall points out that in Gilbert's view, "two individuals can agree to do something

together […] only through interactions that are at least implicitly second-personal", and Darwall concludes that "the capacity of individuals […] to form plural subjects depends upon their already presupposing one another's second-personal standing in seriously addressing each other in the first place" (202).

If that is all there is to Darwall's slogan "the way to the 'we' leads through 'you' and 'I'", I have nothing to criticize. Gilbertian plural subjects really do presuppose the structure of mutual recognition Darwall has in mind, however tacit or implicit the agreement in question may be. However, it seems to me that Darwall is up to more than that, at least in parts of his book. There are two potentially stronger readings, neither of which is explicitly endorsed, as far as I can see, but at least one of which seems to be operative in Darwall's interpretation of the Milgram experiment. The first stronger reading amounts to the claim that the 'we'—at least the 'we' that is relevant for forms of cooperation that involve normative claims—is *nothing but* some structure of mutual second-personal attitudes, so that an analysis of the "Second-Person Standpoint" yields an understanding of what it means to share a standpoint and to cooperate. According to this view, it is not just the case that the 'we' *presupposes* some 'you' and 'I'; rather, 'you' and 'I' is *all there is* to the 'we'. This leads to a problematic view of how interpersonal relations work in cooperation. But such an eliminative reductivism seems to be alien to Darwall and contrary to his claim that the second-person viewpoint is "a variant" of the first-person plural standpoint. Another reading—still a rather strong one—may be at the back of Darwall's mind; it is the claim that the sort of "we" that is in play whenever the actions and intentions of the participants must be seen as contributions to a shared single whole in order to be understood correctly is really based on a different "variant" of the first person plural, which in turn is really "nothing but" second-personal. But there is no level at which the second-person viewpoint does not involve a sense of "us". The case of the interpretation of Milgram's experiment illustrates vividly what it means to miss the "we" for all the "you's" in that way. In cooperation, any individual action and intention receives its meaning from its context; this is true for address and authority. The degree to which we are "bound" by each other's claims, and subject to each other's authority, depends on what it is we take ourselves to be doing together. And doing things together may not be the only form of human sociality, but it is certainly important.

References

Bierbrauer, Günter 1979: "Why Did He Do It? Attribution of Obedience and the Phenomenon of Dispositional Bias". *European Journal of Social Psychology* 9, 67–84.

Blass, Thomas 2004: *The Man who Shocked the World. The Life and Legacy of Stanley Milgram.* New York: Basic Books.

Clarke, Herbert H. 2006: "Social Actions, Social Commitments". In: Nicholas J. Envield et al. (eds.), *Roots of Human Sociality.* Oxford: Berg, 126–149.

Darwall, Stephen 2006: *The Second-Person Standpoint. Morality, Respect, and Accountability.* Cambridge, MA: Harvard University Press.

Davidson, Donald 2001: *Essays on Actions and Events.* Oxford: Oxford University Press.

Fromm, Erich 1973: *The Anatomy of Human Destructiveness.* New York: Holt et al.

Gibbard, Allan 1990: *Wise Choices, Apt Feelings. A Theory of Normative Judgment.* Oxford: Clarendon Press.

Gilbert, Margaret 2010: "The Practical Import of Commands". Unpublished Manuscript.

Milgram, Stanley 1974: *Obedience to Authority. An Experimental View.* New York: Harper.

Railton, Peter 2006: "Normative Guidance". In: Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics,* Vol. 1. Oxford: Oxford University Press, 3–34.

Schmid, Hans Bernhard 2011: *Moralische Integrität. Kritik eines Konstrukts.* Berlin: Suhrkamp Verlag.

Schmid, Hans Bernhard & Schweikard, David 2013: "Collective Intentionality". In: *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2013/entries/collective-intentionality/>.

Solomon, Robert C. 2003: "Victims of Circumstances? A Defense of Virtue Ethics in Business". *Business Ethics Quarterly* 13/1, 43–62.

Grazer Philosophische Studien
90 (2014), 63–77.

DEMANDING SOMETHING

Peter SCHABER
University of Zurich

Summary
Stephen Darwall thinks that moral obligations are second-personal, demands that are addressed to others. But what is it that is addressed to others when we demand of them certain things? It will be argued that there are different forms of demanding things. Some but not all demands have normative force. None of these forms of demands, however, can help us to understand Darwall's idea of moral obligations as second-personal demands. Rather it remains unclear how moral demands could be conceived of as second-personal addresses.

It is a central idea of Stephen Darwall's second-personal ethics that there is a conceptual link between moral obligations and demands we can address to each other. Moral obligations, he puts it, "conceptually involve addressable … authoritative demands. When we violate moral obligations, we violate legitimate expectations and demands that we have and can make …" (Darwall 2013, 61). If a person is under a moral obligation to do x, doing x can be demanded of her. Moral obligations, Darwall thinks, do not have to be demanded to be obligations. They "do not depend on being made by anyone with the individual authority to make them" (Darwall 2013, 35). But they can be demanded. And they can be demanded by those to whom the person who has the moral obligation is accountable. According to Darwall, obligations are genuine demands.

What does this tell us about the nature of moral obligations? Darwall thinks that moral obligations are second-personal, demands that are addressed to others. But what is it that is addressed to others when we demand of them certain things? That is: What are demands? This question needs to be answered if we want to get a grip on the idea that moral obligations have to be conceived of as genuine demands. Darwall talks of "valid demands". They presuppose, as he thinks, "the authority to make it and that the duly authorized claim creates a distinctive reason for com-

pliance (a second-personal reason)" (Darwall 2006, 11). This authority also involves the authority to hold the other person responsible: If you're authorized to demand of someone to do x, you are also authorized to hold her or him responsible (see Darwall 2006, 11). This is the way Darwall conceives of demands and this is also all he says about the concept. This does not answer the question of what we are doing when we are demanding things. I will argue that there are different forms of demanding things. None of them can serve, I will argue, as a plausible explanation of what it means to be under a moral obligation.

I will argue that some, but not all demands have normative force. Those that have normative force do change the normative properties of a situation. Some of them, it will be argued, create moral obligations; others just change the reasons for action people have by putting pressure on these people. Some demands have no normative force at all. They just make the normative set up of a situation explicit. They, for instance, are just making it clear to other persons that they are under a moral obligation that obtains independently of having been made explicit. None of these forms of demands, it will be argued, can help us to understand the idea of moral obligation Darwall has in view when he talks of moral obligations that do not depend on being demanded by anyone with the individual authority to make them. Most moral obligations belong to this category of obligations. One ought not, for instance, kill, torture, degrade, or betray others regardless of whether this is demanded by the person concerned. Darwall also thinks that we ought to do certain things regardless of whether these things are actually demanded of us. But the question is whether this view is compatible with Darwall's view, according to which legitimate moral demands are *second-personal*. I will argue that this is not the case, because only actual demands create second-personal reasons. Moral demands are, as also Darwall agrees, not actual demands. But if this is the case, it is not clear how moral demands could be conceived of as second-personal addresses. This is the point I want to make in this paper.

1. *The normative force of demands*

Let us first have a closer look at Darwall's understanding of moral demands. Take Darwall's example: I step on your feet. I fail to appreciate the reasons not to do so. You could demand of me to take my foot off your foot: "I'm sorry, but that's my foot you're stepping on, you might say to me …"

(Darwall 2013b, 67). You would address me second-personally, by making a claim on me. Demanding is addressing yourself to another person. Doing so creates second-personal reasons. Thus, according to Darwall, addressing yourself to others has normative force.

> When a sergeant orders her platoon to fall in, her charges normally take it that the reason she thereby gives them derives entirely from her authority to address demands to them and their responsibility to comply … The sergeant's order addresses a reason that would not exist but for her authority to address it through her command. (Darwall 2006, 12pp.)

The reason to comply has been created by the sergeant's command. "Similarly", Darwall adds, "when you demand that someone move his foot from on top of yours, you presuppose an irreducibly second-personal standing to address this second-person reason" (Darwall 2006, 13). I step on your feet. I ought to take off my foot, and I'm obliged to do so, as Darwall thinks, for various reasons. There is first the pain I cause; but then, in addition, there is your legitimate demand that I take off my foot. That is, I ought to take it off also because you *can* demand this of me. I do not have this obligation because you actually demand it of me, rather because you could rightly demand it of me. Darwall's idea is that legitimate demands are providing people with reasons for action. They do this in addition to the other reasons persons have to do certain things. But it is the demand itself that binds me to take off my foot, because doing so "violates your right, that is, on the current analysis, your legitimate demand of me as an individual person" (Darwall 2013b, 67).

It is important to note that your *actual* claim on me ("I'm sorry but that's my foot") has no normative force. It is not the case that I ought to take off my foot *because* you addressed me second-personally. I would be obliged to do so even if you said nothing. Thus, it is not the actual act of demanding that carries the normative force in this case, it is rather the *possible legitimate* act of demanding that provides me with additional reason. Darwall understands demands as moral rights. Moral rights are things that can be claimed, but they do not have to be claimed to be rights that have to be respected by others. Rights in the moral sense are, as Darwall thinks, legitimate demands that can be claimed by those who have them. "Our moral rights are what we have the authority to demand of one another *as individuals*" (Darwall 2013, 67).

2. *Demands and requests*

What exactly does it mean to demand of others certain things? Darwall thinks that by demanding we are addressing claims to each other second-personally ("I'm sorry, but that's my foot"). There are ways of addressing claims to each other, other than demanding. We also address claims to others by requesting things. "Could you please help me writing this paper?", a colleague could ask me. Requests are second-personal: they are addressed to others. They share this feature with demands. What then distinguishes requests from demands?

John Searle thinks that they just differ in strength (see Searle 1976, 1–23). They are all attempts to get others to do certain things. Requests according to Searle put less pressure on others than demands in order to reach the envisaged end. However, I think that demands differ from requests not by the degree of pressure they put on others. Requests and demands are different ways of addressing yourself to others second-personally. Requests and demands have, as Lance and Kukla argue, different "normative outputs": "The output of a successful imperative is an obligation … The output of a successful request is that the target now has a specific sort of reason to do what was requested …" (Lance/Kukla 2013, 460).[1] Requests create reasons for doing something that remains to be an option, not an obligation. A request according to this would be: "It would be nice of you if you helped me writing my paper, but you do not have to do this". Requests, then, are normatively not neutral by creating reasons for action. If the person the request has been addressed to does not respond to these reasons, she does not act the way she should. But she cannot be blamed, because requests do not create obligations. Requests grant options: The reasons they are providing people with are not silencing the other reasons they have. My colleague has reasons to help me writing my paper, but she has at the same time reasons to do other things. Demands, on the contrary, do not grant options. If I demand of you to take your foot off my foot, you do not have options. To take off your foot is something you must do.

Demands bring about obligations, requests reasons for action, both however only if they are successful. When are they successful? In both cases one must have the authority to request or to demand something. Only if I have the authority to ask the target person to help me finishing

1. Lance and Kukla talk of imperatives and not of demands. I take them to have the same meaning.

my paper, my request creates reasons for action for her. I do not have, for instance, the authority to ask the vice-chancellor of our university to help me—if I asked him, this would not create reasons for him to do so. Of course, he might have independent reasons to help me (he wants his colleagues to be academically successful). But my request does not create additional reason for him to act accordingly. Our relation is, I assume, not such that I could address him second-personally in this way. I do not have the authority to make successful requests. The same holds for demands. One must have the authority to demand something of someone to have an obligation as a normative output.

And when requests or demands are successful, different reactive attitudes are warranted when the target persons do not act the way they were supposed to act. If the other person does not do what I requested her to do, I could rightly be disappointed. She does not seem to be the good colleague I thought she was. This might not just change my view of her, but also my relation to her. If the other person does not do what can legitimately demanded of her, I'm justified in blaming her. I might, of course, also be disappointed; but I'm in addition to this justified in putting blame on her. She has not done what to do she had an obligation. She can be blamed for that, which is not true in the case of a request.

3. *The act of demanding*

Demands, if successful, have (at least in certain cases) a certain normative output. But what are we doing when we demand something of others? This is still unclear. One could believe that acts of demanding are attempts to put other people under obligations, as requests might be understood as attempts to provide others with certain reasons for action. Take the case of a thief who wants me to hand over my money. He does not want to grant me further options ("it would be nice if you gave me your money, but you don't have to"), he wants to make it a must for me to do so. So he threatens to kill me if I don't hand over the cash: "Now give it to me". He intends to leave me without options; still he does not intend to put me under a moral obligation. He would not blame me if I refused to give him my money. He would be angry and would possibly kill me, but he would not think that what I did was morally wrong, a violation of a moral obligation. Thus, demanding is not necessarily an attempt to put another person under a moral obligation.

But what are we doing then when we demand something of others in the moral sense? Demanding is, as Darwall emphasizes, conceptually linked to blaming. Thus, we have demanded something of others, one could argue, if we blamed her if she did not do what we demanded of her. "I'm sorry, but that's my foot you're stepping on", A tells B. Is this a demand or just a request? Does this depend on whether A will blame B if he refuses to take off his foot? Can we hence only know whether it is a demand when we know whether or not A is going to blame B in case of B's non-compliance? But blaming is in any case not what A is doing when she demands of B to take off her foot. A is demanding something of B simply by saying, "Take off your foot". What then turns this into a demand?

Is it the pressure on the target person that turns the utterance into a demand? A could threaten B by raising his arm or other gestures. But there are demands that are without doubts demands where no pressure is exercised. The head of department might write me a letter telling me to replace a colleague as an examiner next week: "I expect you to act as examiner next Tuesday at 2 p.m., room 214". She does not threaten in case of defiance, she just tells me to act as an examiner. And this puts me under an obligation, provided she has the authority to demand this of me. I could be blamed by her, and I might even be sanctioned if I did not act as an examiner without having a good excuse.[2]

What is going on here? Provided that what the head of department tells me is meant as something I must comply with, it is an act of demanding. Demanding in this case is not just creating reasons for me to act, it is putting me under an obligation, which I did not have if the head of department did not demand of me to act as an examiner. I think that the head of department creates this obligation by exercising a normative power she has over me. She has the right to tell me that I have to examine philosophy students, because she has the duty to organize the examinations at our department. Due to the latter, she has the authority to demand of me to act as an examiner. And because she has this authority, she is able to put me under the moral obligation to act as an examiner by telling me to do so. It is her authority that puts me under an obligation. More precisely, it is her exercise of the normative power she has over me as head of department ("I expect you to act as an examiner…") that puts me under an obligation. Thus, demanding can be understood as exercising the norma-

2. It is not necessarily the case that she is justified in sanctioning me. This is at least not implied by being under an obligation.

tive power one has over others. This is also what the sergeant does when he demands of one of his soldiers to dig a hole or what the chief editor of a newspaper does when he tells one of the journalists to write an article on the Arab revolution.

When, however, the thief wants my money, he is not exercising a normative power. If what he does is nevertheless demanding, it is a different form of demanding. Unlike the forms of demanding we have considered above, it has no normative force. It is neither providing others with reasons for action nor is it putting them under an obligation. The thief behaves as if he had the required normative power over me. But because he does not have it, he is not successful in putting me under an obligation. He might, however, be successful in another way, namely in actually getting my money. This is not due to his demand; this is rather due to his putting me under pressure by threatening me. So we have two different forms of demanding: the one which is the exercise of a normative power and the other which is just putting pressure on others. The former creates obligations; the latter fails to do so or does not even intend to do so. Thus, demands come in many varieties. They differ in particular with regard to their normative force: some have normative force, others do not. A demand is not necessarily the exercise of a normative power.

4. *Making it explicit*

Let us come back to Darwall's claim that moral obligations are legitimate moral demands. The question is whether the things we've just said about demands are compatible with Darwall's understanding of moral obligations as demands. Darwall thinks that people have the authority to demand of each other certain things. And he holds that "no moral obligation period can exist unless *non-discretionary* demands exist that do not depend on being made by anyone with the individual authority to make them or not" (Darwall 2013, 35). Darwall's idea is that there could be no moral obligation if there were no legitimate demands to be made. Is a legitimate moral demand understood this way as an exercise of a normative power?

Let us have a look again at the "stepping on your foot" example. I step on your foot. I have an obligation to take my foot off your foot. You are, of course, entitled to demand of me to take off my foot. But your saying "I'm sorry, but that's my foot you're stepping on" would not put me under

an obligation to take off my foot. I'm obliged to do so, independently of your saying this to me.

Demanding as an exercise of a normative power, in contrast, changes the normative property of a situation. I ought to act as an examiner if the head of department demands of me to do so; and I would *not* have to do if she did not demand this of me. It is the exercise of her normative power over me that brings about this normative change, from something I'm not obliged to do to something I'm now obliged to do. Your demand to take off my foot, in contrast, is not an exercise of normative power, because it does not change the normative properties of the situation. I'm obliged to take off my foot, even if you said nothing. By saying so you are reminding me of having this moral obligation. Your demand is legitimate, but it does not change the normative properties of the given situation.

You could change the normative properties this way: You could say, "Stepping on my foot is fine with me; you do not have to take your foot off". If you did this, you would waive your right not to be stepped on your foot. This is not demanding anything of me. On the contrary, this is the waiving of a demand you are entitled to make. Demanding of me to take off my foot is neither the exercise of a normative power nor is it a failed attempt to exercise a normative power. That is, it is not the case that you aim at a normative output you do not succeed in bringing about. The latter holds when you did not have the normative power to demand of me to take off my foot. But you do have the normative power to waive your right not to be stepped on. You are entitled to demand this of me, but your doing so does not create a reason to comply with the obligation. The normative properties of the situation are not subject to any demands, only to acts of waiving your entitlements. If so, the demands are not second-personal the way Darwall understands demands as being second-personal. As he puts it:

> A second-personal reason is one whose validity depends on presupposed authority and accountability relations between persons and, therefore, on the possibility of the reason's being addressed person-to-person. Reasons addressed or presupposed in orders, requests, claims, reproaches, complaints, demands, promises, contracts, giving of consent, and so on are all second-personal in this sense. (Darwall 2006, 8)

The validity of the reasons the demands of the head of department provide me depend on the exercise of her normative authority. As concerns moral obligations, this is not the case. The reasons one has to comply with them

are prior to the demands and also prior to the authority to demand. They are the reasons why people have this authority in the moral case.

5. *Demands and obligations*

Obligations *can* be the normative output of demands: of exercises of normative powers people have. They presuppose that certain normative relations obtain between people: A has the normative power over B's doing x: A's demand of B to do x creates an obligation on B's side. Most moral obligations we have, however, are not of this kind: The duty not to cause pain or the duty not to kill are simply there to be fulfilled. People are entitled to demand of others not to cause pain etc., provided the others are obliged not to do so. There, hence, is a difference between demands creating obligations and demands that make obligations explicit. The former have normative force, the latter are normatively inert.

Darwall thinks that there could be no moral obligations period "unless non-discretionary demands exist that do not depend on being made by anyone with the individual authority to make them or not" (Darwall 2013, 32). What Darwall seems to mean here is this: A is morally obliged to do x means that others are authorized to demand of A to do x. If they were not authorized to do so, A would not be obliged to do x. But does being obliged to do x mean that doing can be demanded by others?

I do not think that this is what follows from what above has been said about demands. The demand in Darwall's account is something that is justified by a moral obligation. What is justified here is a demand understood as an utterance that makes the moral obligation explicit. That is to say, one is right in saying, "you ought to take off your foot" because the other person is under a moral obligation to do so. The moral obligation is logically prior to any act of making it explicit. It is simply what is made explicit by the demand. This is a problem for Darwall's account of moral obligations. It is unclear how moral obligations could be conceived of as second-personal. This is so because moral obligations do not depend on the demand. On the contrary, the legitimate moral demand presupposes the existence of an independent moral obligation. This is the case where demands are acts of making it explicit. It is different, however, from demands that are exercises of normative powers: these demands create moral obligations. My demand creates an obligation, provided I have the required normative power.

These two kinds of demands differ not only with regard to their normative force, they also differ in another respect: If I step on your feet, every one is authorized to demand of me to take off my foot, not only the person who suffers. If she did not say anything, the person sitting next to her could say: "Please take off your foot, can't you see what you're doing to her?" But if I have to write an article just because the head of department asked me to do so, only my boss would be authorized to demand it of me in case I do not follow his demand. It would be none of anyone else's business to demand this of me. Of course, others could tell me that the head of department asked me to write the article, but they were not in a normative position to demand of me to do so. The authority to demand of me to act appropriately would remain with the person who brought the obligation into existence: the head of department. But in the moral case every one was entitled to demand of me to take off my foot, because the reasons to do so exist independently of any exercises of the normative authority people have.

6. *Second-personal?*

Darwall thinks that moral obligations conceptually involve addressable demands. As we've seen, some moral obligations are the result of successfully addressed demands: those that are brought about by exercises of normative powers people have. Most moral obligations are not of that kind. They exist independently of whether they are demanded or not. They could be demanded in the sense that they could be made explicit. But this is not part of what it means to be under a moral obligation. Various things can be made explicit. Everything that is the case, for instance, can be made explicit. I can tell others that the earth is round; in the same way I can tell them that they ought not to humiliate their fellow human beings. We would not want to say that nothing was the case unless it could be made explicit. In the same vein, we would not say that there was no moral obligation unless there were addressable demands. That they can be demanded presupposes that they exist independently of whether they are addressable to others.

Moral obligation involves demands that are addressed to us by the performance of a particular act on a particular occasion. Of course, they can be addressed to us in this manner, but they are not created by a particular act. They are addressed to us, as Darwall puts it, "from the … perspective

of the moral community" (Darwall 2006, 9). And if others address moral demands to me, they do this as representatives of the moral community. What you are obliged to do is what the moral community demands of you. Are the demands issued by the moral community just making my moral obligations explicit or are the moral obligations created by them? Darwall's idea seems to be that the community's demands put me under moral obligations. This presupposes that the moral community has the required moral power the exercises of which create moral obligations. I have to take off my foot because the moral community demands me to do so.

But this is not an act that is performed by real people in the real world. Thus, it is not the exercise of a normative power. The exercise of normative power is an act that is performed by someone. It is an act that the one who has the required normative authority could perform. But this does not apply to the demands of the moral community simply because they are not actually performed. Of course, representatives of the moral community would demand of you what morality requires. But it seems that by doing this they were making explicit the moral obligations we have to fulfil. Exercises of normative powers as acts of demanding in the sense described above are acts that are performed by particular people on particular occasions. What the moral community on Darwall's view does is not of this nature. Thus, their demands are not creating moral obligations. The moral obligations are presupposed by moral demands. They must be conceived of independently of moral demands because they do not depend on the normative authority people have. Darwall writes:

> The sergeant's order addresses a reason that would not exist but for her authority to address it through her command. Similarly, when you demand that someone move his foot from on top of yours, you presuppose an irreducibly second-personal standing to address this second-personal reason. (Darwall 2006, 13)

But the two cases are not similar as concerns the reasons for action. In the first case the reason would not exist without the command, in the second, however, it would. This is the difference for which Darwall does not account.

7. *Explaining the normativity of obligations*

Let me look more closely at the obligations that exist independently of demands. This is important because Darwall seems not to accept such demand-independent obligations. In order to show that Darwall should accept such demands, we need to sketch how these demand-independent obligations might come into existence.

First of all, Darwall is right, I think, that the concept of moral obligatoriness differs from simply having reasons (see Darwall 2013a, 33). I do not just have reasons to take off my foot from your foot: I'm obliged to do so. This is also a difference in how the others could relate to my taking off my foot: It is not just that they could recommend to do so, they can demand this of me. And because this can be demanded, I'm blameworthy if I don't take off my foot. Those who can demand this of me can blame me for not doing it. Darwall rightly thinks that all who have the representative moral authority can do so.

> (W)hen we blame someone, we add our voice to or second, as it were, a demand that we must presuppose is made of everyone by the moral community or representative persons as such. (Darwall 2013a, 37)

We might criticize people for not paying attention to reasons, but we don't blame them. We blame them only when we think that they violated a moral obligation. Obligations have a binding force. If I'm under an obligation to do x, doing x is not just an option I can choose, it is something I must do.

How can we account for this difference? An option is choiceworthy to the extent that there are reasons to choose the option. There are reasons to do what is obligatory. But if I'm obliged to do x, x is not just something I have reason to do. What then has to be added when it comes to actions that are obligatory?

Take the Darwall's toe case again. Stepping on your foot hurts you and is therefore bad for you. This is a reason not to step on your foot. On Darwall's view there is an additional reason not to perform such acts:

> (A)n act's wrongness provides additional reason not to perform it: the act would violate a legitimate demand and so fail to respect our authority as representative persons. (Darwall 2013, 69)

But why is this something that can be demanded of others? I have a duty not to step on other's feet. They can demand of me not to perform such acts, because it is my duty not to perform them. This is not due to your

demand. It is not the case that your demand creates my duty not to step on your foot. It is my duty not to do so, even if you did not demand of me not to do it. Your demand is just making my duty explicit.

What makes it the case that you have the authority to demand of me not to perform such acts? It's bad for you and thus you have a reason not to want it. Having a reason not to want others to step on your foot does not imply that I have a duty not to perform this act. We have reasons not to do certain things without necessarily having a duty not to do so. For instance, we have reasons to be nice to people without having a duty to be nice; this cannot be demanded by others. But we have a duty not to step on other's feet. Why?

We do not want others to step on our feet. We also want to be able to make sure that others do not perform such acts. We do not want it to be the case that we can only hope that others do not perform such acts. We could think of a world where people could just hope and complain afterwards when others do not behave the way we want them to. We could ask others not to step on our feet. But I think we want to be in a stronger position towards others when it comes to them causing us pain. We have an interest not to be stepped on our feet and in addition an interest in having the normative power to put pressure on others to refrain from doing so. Such an authority is of value to us. It is important to us not to be hurt by others and being authorized to make sure that they will not. We would not want to live in a world where we just have to hope not to be hurt or where we could only ask for not being hurt by others. We want to live in a world where we are authorized to put pressure on others not to perform acts that hurt us.

We are authorized, one could argue, to put pressure on others not to perform such acts, because it is of *value* to all of us to have this normative power. The value of having normative power over certain acts could explain the difference between just having reasons not to perform such acts and being under an obligation not to perform them, a difference Darwall himself thinks does obtain. But is the view according to which normative powers are value-dependent compatible with the idea of moral obligations as second-personal? Darwall writes:

> I argue that the second-personal character of central moral concepts has fundamental implications for the kinds of reasons it takes to justify beliefs and attitudes that involve these concepts … I maintain that it does follow from my analysis … that there is a fundamental conceptual difference between the good and the right, and that considerations showing that an action would be

> desirable … are not the "reasons of the right kind" to establish by themselves the action's deontic status, its being either morally obligatory, prohibited, or permissible. (Darwall 2013, xi)

It is desirable not to step on others feet. Darwall is, I think, right, that this does not put us under a moral duty not to perform such acts. But then it is desirable to have the authority to put other people under pressure not to step on our feet. This is the reason, I guess, to see people under a duty not to perform such acts. It is of value to us that we are authorized to tell others that they are under an obligation to perform certain acts. It is good for us that people do not behave in certain ways. It is also good for us that we can treat them as being obliged to act in these ways. It might be this value that establishes the deontic status of acts; for instance, the deontic status of the act not to step on other people's feet as morally obligatory. We do not want to live in a world where we can only hope that this is not done to us, or where we can only ask others not to treat us this way. We want to live in a world where we can treat people as being under an obligation not to hurt us.

8. *Conclusion*

Darwall thinks that moral obligations are genuine demands we can address to each other. Moral obligations can be demanded by others, without depending on being demanded. But as it has been argued in this paper, obligations are either created by demands or presupposed by them. Demands can have different forms. Demands can be ways of putting others under pressure to do certain things or exercises of normative powers people have. Exercises of normative powers create obligations. These obligations do not exist independently of the people's exercises of their normative power. They are their normative output. Other demands just make obligations explicit. These do not create obligations. Many moral obligations we have, such as the obligation not to kill, not to torture, not to degrade people, are just made explicit by demands. We have these moral obligations not because they are demanded, but rather because we have an interest in taking each other as having such moral obligations. That is to say, we have an interest in being able to bind others to treat us in certain ways, an ability we would not have if we had no right to tell them to be obliged to act in ways they should treat us.

References

Darwall, Stephen 2006: *The Second-Person Standpoint.* Cambridge, MA: Harvard University Press.

— 2013a: "Bipolar Obligation". In: Stephen Darwall (ed.), *Morality, Authority & Law.* E*ssays in Second-Personal Ethics I*. Oxford: Oxford University Press, 20–39.

— 2013b: "But It Would Be Wrong". In: Stephen Darwall (ed.), *Morality, Authority & Law. Essays in Second-Personal Ethics I*. Oxford: Oxford University Press, 52–72.

Lance, Mark & Kukla, Rebecca 2013: "Leave the Gun: Take the Cannoli! The Pragmatic Topography of Second-Person Calls". *Ethics* 123/3, 456–478.

Searle, John R. 1976: "A Classification of Illocutionary Acts". *Language in Society* 5, 1–23.

Grazer Philosophische Studien
90 (2014), 79–104.

REACTIVE ATTITUDES, DISDAIN AND THE SECOND-PERSON STANDPOINT[1]

Alexandra COUTO
University of Oslo

Summary

Disdain and contempt have often been taken to be vicious attitudes. This view has been further defended by Stephen Darwall in his ambitious and elaborate second-personal account of morality. In *The Second-Person Standpoint*, Darwall argues that disdain is problematic to the extent that it fails to recognize the authority and moral freedom of its object. In this paper, I will develop two answers to Darwall's claims about disdain. First, I will argue that, if we take Darwall's account to be ultimately grounded on what hypothetical members of the moral community would do, it would be difficult to argue that anything said at this level would justify an evaluation of what actual individuals should do. Second, even if we granted that the *Second-Personal Standpoint* could have such normative implications, I will argue that disdain, as a moralised attitude towards others who fail to behave morally, can not only be justified but can also be shown to presuppose the moral freedom of the wrongdoer. Finally, I will introduce a distinction between normative and empirical expectation to further clarify this point.

1. *Introduction*

In this paper, I will argue that, contrary to what Darwall has suggested, the Second-Person Standpoint cannot offer theoretical support for the view that we should not feel disdain and contempt towards wrongdoers. On the contrary, I will argue that it is not necessarily morally problematic to disengage from feeling resentment and anger towards those individuals who are prone to commit wrongdoings, and that feeling disdain or contempt, both attitudes of disengagement, therefore need not be morally problematic.

1. Work on this paper was supported by funding from the European Research Council and the Research Council of Norway.

I will start by giving a brief description of the general account defended by Darwall in *The Second-Person Standpoint* (henceforth SPS). I will discuss in particular the role of the reactive attitudes in this account and their relationship to other crucial concepts in the SPS, such as second-personal competence, second-personal reasons and moral obligation. I will then examine the claim made by Strawson and endorsed by Darwall that the reactive attitudes constitute a form of *demand* for respect.

If reactive attitudes issue a demand, giving up on this demand by experiencing disdain could be problematic in various ways. I will discuss three ways in which experiencing disdain or contempt can be said to be problematic in this regard. First, it could imply that the victim gives up, so to speak, on having her demand be answered and this could indicate lack of self-respect. Second, failing to experience the reactive attitudes and experiencing disdain instead might suggest that no moral obligation has been violated. Third, experiencing disdain could indicate lack of respect towards the wrongdoer. Darwall himself endorses this last view. In some passages of the SPS, he describes disdain towards the wrongdoer as failing to show respect to the wrongdoer.

In this paper, I will counter these views by defending the admittedly modest claim that contempt or disdain is not necessarily morally problematic. For this purpose, I will develop two arguments. To begin with, I will point to a general ambivalence in the SPS between an account rooted in the *actual* experiences of individuals and an argument about what dispositions *ideal* members of the moral community would have. If reactive attitudes are understood to be a form of demand for respect, as Darwall argues, then it matters a great deal whether this demand is formulated when individuals are *actually* experiencing the reactive attitudes or whether this demand exists in virtue of the disposition of *ideal* members of the moral community. I will argue that ultimately the argument as presented by Darwall is best understood as concerned with the dispositions of ideal members of the moral community towards the reactive attitudes. To the extent that this is the case, if individuals fail to experience the reactive attitudes and experience instead an attitude of disengagement, such as disdain or contempt, this doesn't undermine the fact that a moral obligation has been violated as only the disposition of ideal members ground moral obligation. The argument presented in this section thus seems to reject only one possible ground for taking disdain and contempt to be morally disvaluable. However, I point out that this argument would also undermine the validity of the two remaining arguments mentioned. This

is because the ideal character of Darwall's account ends up blocking it from having any normative implications.

But, given that Darwall could still argue that we ought nevertheless to strive to emulate the dispositions that hypothetical members of the moral community would have, I will explore another ground to deny the claim explicitly held by Darwall, that is, the claim that experiencing disdain might fail to recognize the respect owed to the wrongdoer as a morally free agent. For that purpose I will draw on the distinction between recognition and appraisal respect introduced by Darwall himself in an earlier work in order to show that feeling disdain towards a wrong-doer *also* presupposes that the wrongdoer belongs to our moral community. Moreover, I will introduce a distinction between empirical and normative expectation in order to argue that although stopping to expect better from particular agents at a *normative* level might be problematic, stopping to expect better at an *empirical* level is in fact often more appropriate. And I will claim that disdain requires only the latter. The argument concludes thus that disdain is not necessarily disvaluable or morally inappropriate.

2. *Darwall's second-person standpoint: morality as equal accountability*

Let me start by giving a brief sketch of the view developed by Darwall in *The Second-Person Standpoint*. Darwall describes the second-person standpoint as the "perspective you and I take up when we make and acknowledge claims on one another's conduct and will" (Darwall 2006, 3). What this perspective is getting at ultimately is a conception of morality as a form of *equal accountability* to each other. This conception of morality as a form of equal accountability emerges from a circle of different concepts all linked together (as Darwall himself describes it): moral obligation, second-personal reason, second-personal competence and the authority to make demands on others.

On Darwall's account, the concept of moral obligation can only be understood from a second-personal perspective, that is, from the perspective of what *you* can demand from *me* and what *I* am accountable to *you* for. Moreover, I can only be morally obligated to do what there is a second-personal reason for me to do. A second-personal reason is a reason that depends conceptually on the victim having authority to make claims or demands. Second-personal reasons also imply an individual being addressed by them. Only second-personal reason can ground claims

of moral obligation and thus rights. This being said, Darwall recognises the existence and relevance of agent-neutral reasons: the fact that if I do something I might cause pain to another being might still have a role in determining what to do. However, on Darwall's view, this being can only have *a right* not to suffer pain caused by me if it has the *authority* to make claims or demands from me, that is, if there are second-personal reasons for me not to cause pain to it. Agent-neutral reasons, unlike second-personal reasons, are not conceptually connected to what we are morally obligated to do. And beneficial consequences pointed to by agent-neutral reasons can't ground any right.

So now it becomes pertinent to wonder what gives us this authority to make claims on others. If second-personal reasons depend on the victim's authority to make claims, what underlies the authority to make claims? Darwall answers this question by pointing to second-personal competence. According to him, second-personal competence makes us subject to moral obligation and also reciprocally gives us the authority to make claims and demands of one another as members of the moral community (Darwall 2006, 29). It is "the moral competence requisite for (equal) membership in the moral community" (Darwall 2006, 75).

But we might now wonder: what then grounds second-personal competence? What grounds second-personal competence is the ability to comply with norms (Darwall 2006, 242). This compliance with norms should be done for the right reasons, that is, because the agent recognises the authority of the potential victims of her wrong-doing to make claims on her. Described in this way, second-personal competence is thus connected to autonomy. Darwall argues indeed that persons can have second-personal competence "only if we can assume autonomy and some form of moral reasoning like the Categorical Imperative." And in virtue of having second-personal competence, one attracts certain privileges: one can claim, "respect for persons" and be acknowledged as free and equal (Darwall 2006, 127).

However, unlike the Categorical Imperative, second-personal competence doesn't merely require a detached form of moral reasoning. This moral reasoning should be accompanied by the appropriate emotional attitude towards other agents. Darwall takes deficit in empathy to be deficit in second-personal competence (Darwall 2006, 75, 25f). Therefore, although high functioning autists can follow certain norms, they don't acknowledge second-personal authority and can't understand the basis for the principles of moral obligations that they are following. Autistic

individuals, along with psychopaths, lack second-personal competence. However, once someone has second-personal competence, the degree to which this person has second-personal competence is irrelevant as it is a range property (Darwall 2006, 75).[2]

To sum up, according to Darwall, in order to understand what is going on in moral accountability, we need to consider a "circle" of mutually involving concepts, which presuppose the second-person standpoint and can't be accessed from outside it. These concepts include the authority to make a claim or demand, the authoritative claim or demand itself, the second-personal reason for complying, and the accountability for complying. I can only have the authority to demand that someone refrains from wronging me if my demand is legitimate, if there is a second-personal reason for this other person to refrain from wronging me, and if she is accountable to me for this wrongdoing. Moral accountability is thus irreducibly second-personal (Darwall 2010, 217–218).

3. *The role of the reactive attitudes in the second-person standpoint*

As mentioned above, Darwall argues that the concept of moral obligation should be understood from a second-personal perspective, that is, from the perspective of what *you* can demand from *me*. More specifically, what constitutes a moral obligation for me is what, from this perspective, would warrant the experience of the *reactive attitudes* (in particular blame and resentment) on your part.

The claim that moral obligation is best understood as what would warrant the experience of the reactive attitudes is a variation on Strawson's claim in his classic article 'Freedom and Resentment' (Strawson 1974). In 'Freedom and Resentment', Strawson claimed that human beings are disposed to experience negative reactive attitudes towards other agents who have wronged them as long as the wrongdoer is a "psychologically normally developed" agent and is acting voluntarily and in full knowledge. On a conceptual interpretation of Strawson's claims, to hold someone responsible implies being disposed to experience the reactive attitudes towards

2. The concept of a range property was introduced by Rawls and stipulates that everyone who possesses a property within the range is equally within the range (Rawls 1999, 443). In other words, nobody is a more competent member of the moral community; once you are a member of the moral community, you are an equal member of the moral community, and nothing more needs to be said about your level of competence.

that person. Darwall has slightly revised Strawson's claim by specifying that "what we are morally obligated to do is, as a conceptual matter … what would warrant blame … if done without adequate excuse" (Darwall 2010, 221). Darwall takes someone to be responsible when she is justifiably held accountable for her conduct and thus rightly subjected to the reactive attitudes of the victim.

If the content of a moral obligation is determined by what would warrant the experience of the reactive attitudes, one might ask: how is second-personal competence related to the warranted experience of reactive attitudes? On this, Darwall follows once again the stance taken by Strawson. On Strawson's view, experiencing the reactive attitudes presupposes that the addressee of your reactive attitudes has the capacity to understand what is demanded from her and act on it (Darwall 2006, 76). In other words, when you experience the reactive attitudes towards another agent, your experience of the reactive attitudes presupposes the second-personal competence of the individual towards whom you experience the reactive attitudes. I have now discussed what the reactive attitudes imply in terms of holding others responsible for their actions. But I haven't yet said anything about what reactive attitudes *do*. On Darwall's view, reactive attitudes issue *demands* to the person they address. On that point, he follows Strawson who also claimed that "… reactive attitudes implicitly address demands. They involve an expectation of and demand *for* certain conduct from one another" (Strawson 1968, 85). Viewing reactive attitudes as a form of demand is not unique to Darwall or Strawson. Jay Wallace takes the reactive attitudes to "reflect the psychological attitude of demanding", and Gary Watson takes them to be "incipient forms of communication which express demands" (Wallace 1994, Watson 2005).

These demands that we make on one another when we enter the second-person standpoint need not be explicit; they can also be implicit, as some reactive attitudes can be unexpressed feelings like resentment and guilt (Darwall 2006, 3). Moral accountability is second-personal to the extent that it presupposes the victim having the authority to make a demand to be treated otherwise, and the reactive attitudes express this implicit demand. Moreover, Darwall specifies what the reactive attitudes demand; he argues that the reactive attitudes actually demand the reestablishment of reciprocally recognizing respect (Darwall 2006, 60). This demand-making feature of the reactive attitudes will be an important premiss for two of the arguments examined critically below.

4. *Disdain and contempt according to the second-person standpoint*

4.1 *Disdain and contempt*

Let me say a bit more about the attitudes of disdain and contempt. Darwall speaks of disdain, but I take these two attitudes to be closely related and, for the purpose of this essay, I will not distinguish sharply between the two.[3] According to the Oxford English Dictionary, "contempt" derives from the old French "contemner", which means to disdain, so the two concepts are etymologically very close. Let me thus focus here mostly on the notion of contempt, which has been the object of much more in depth analysis. The first thing that ought to be pointed out about contempt is that it is "characterized as a way of negatively and comparatively regarding someone who is presented as falling below the contemnor's personal baseline." (Bell 2013). In other words, the person who is in position of holding someone else in contempt expresses a belief in her own relative superiority. Moreover, this superiority is affirmed when the person held in contempt has failed to meet an interpersonal standard crucial in the contemnor's value system (Bell 2013). This interpersonal standard constitutes the contemnor's personal baseline. When someone stoops below this personal baseline, he/she will be held in contempt by the agent in question.

Another important feature of contempt is that it is a globalist attitude, that is an attitude that takes whole persons as its object (Bell 2013) instead of taking an action as its object (Mason 2003, Bell 2005). Although this characteristic of contempt has been challenged (Abramson 2009, Ben Ze'ev 2001), I will assume here that contempt *is* a globalist attitude. It would be, of course, easier to make a case in favour of a reappraisal of contempt if it were *not* a globalist attitude. But I don't want the defence of the attitude of contempt I will put forward here to be contingent on it not being a globalist attitude. Moreover, one of the arguments I examine assumes also that contempt is a globalist attitude (namely the argument proposed by Darwall that contempt treats the wrongdoer disrespectfully).

Another noteworthy feature of disdain and contempt is that their negative evaluation is coupled with an affective as well as a behavioural withdrawal from and avoidance of the object of contempt and disdain (Bell

3. Scorn might be another very closely connected concept.

2005, 84). This psychological withdrawal is key in differentiating sharply disdain and contempt from other emotional attitudes that negatively evaluate but keep engaging emotionally and behaviourally with the individual in question, such as anger and resentment. When one holds someone in contempt, the object of our contempt doesn't elicit any strong emotional reaction any more and his/her company is no longer sought. In other words, avoidance and withdrawal are essential features of the attitudes of disdain and contempt (Abramson 2009, 202). These features will be crucial in my argument in this paper. I will argue that these features are not necessarily problematic and that, instead, we should take them to be crucial elements in our prerogative as moral agents to decide which interpersonal relations we want to have and which we want to let go.

4.2 *Which normative implications for disdain could be drawn from the second-person standpoint?*

In section 3, we have established that, on Darwall's view as developed in the SPS, the experience of the reactive attitudes is a form of demand for *whatever it is* that would re-establish reciprocal recognition respect, that is, the respect owed to all individuals *qua* morally free agents. This clearly implies that reactive attitudes are valuable and ought not to be abandoned. Disdain (or contempt) however is typically an attitude of disengagement, which is incompatible with experiencing the reactive attitudes. If reactive attitudes involve a demand for respect, the absence of reactive attitudes and the substitution of attitudes of disengagement such as disdain and contempt might appear to be disvaluable. So the disvalue of the attitudes of disdain and contempt seem to be a normative implication of the SPS. There are three different reasons why giving up on the reactive attitudes to endorse the attitudes of contempt and disdain might be disvaluable: it might be seen as amounting to lacking self-respect, disrespecting the object of contempt or undermining the claim that a moral obligation has been violated. Let me say a few words about each of these in turn before I develop responses to these arguments in section 6.

A. *Self-respect*

One reason for taking disdain and contempt to be disvaluable is that the experience of disdain or contempt would amount to giving up on demanding to be treated respectfully. Recall that, if one takes disdain, as Darwall does, to be contrasted with the reactive attitudes, the experience

of disdain replaces the experience of the reactive attitudes.[4] Recall also that, on Darwall's account, the reactive attitudes are a demand for respect. If disdain is taken to replace the experience of reactive attitudes, then it is problematic in so far as no reactive attitude is experienced and no demand to be treated respectfully is made. So if a victim resorts to disdain instead of the reactive attitudes, she is giving up on her demand to be treated respectfully. In support of this view, consider Jeffrie Murphy's claim that:

> not to have what Peter Strawson calls the reactive attitude of resentment when our rights are violated is to convey—emotionally—either that we do not think we have rights or that we do not take our rights very seriously. (Murphy 1982, 17)

According to Murphy, the absence of reactive attitudes in the face of a wrongdoing implies that the victim is not taking her rights seriously and possibly showing lack of self-respect. Murphy is here concerned with the specific reactive attitude of resentment, but one can safely assume that it can be extended to other reactive attitudes, such as anger. Moreover, this view seems to be a normative implication of adopting Strawson's view on the reactive attitudes. I won't discuss specifically this argument but one of the responses I give below (see section 6.1) can be used to briefly address this concern and reject this argument.

B. *Lacking respect towards wrongdoers*

Darwall also argues explicitly that disdain is disvaluable (and seems to suggest that this is a normative implication of the SPS). However, he does so for another reason. The problem, according to him, is not that the disdaining party would show a lack of self-respect towards himself by giving up the reactive attitudes, but that he would be lacking respect towards the disdained party. Similarly, Kant claimed that we had a duty to avoid contempt, because it is incompatible with respect (Kant 1991, 255). In other words, although Darwall also claims that a normative implication of the SPS is that disdain is disvaluable, his account of why and how it is so is wholly different from the one mentioned above.

This view follows from the account presented by Darwall. I mentioned above that a distinctive feature of disdain (and contempt) is its being an attitude that shows behavioural disengagement. This behavioural disen-

4. I will be assuming here and throughout that if one feels disdain or contempt towards some wrongdoer, that attitude normally leaves no space to also feel anger or resentment towards the same person, at least not for the same wrongdoing.

gagement can be taken to show that we are not expecting a change for the better on the part of the wrongdoer. If we take the Strawsonian account seriously, not expecting a change for the better amounts to giving up on the wrongdoer as a member of our community of moral equals (Hill 2000). Given this understanding of these attitudes, they might indeed seem problematic on the second-personal account of morality developed in the SPS.

When Darwall describes the difference between the experience of the reactive attitudes and the experience of disdain, the worry he mainly points out with the attitude of disdain is that it fails to treat the wrongdoer as still belonging to the moral community of equals. According to him, unlike the experience of the reactive attitudes, experiencing disdain writes off the wrong-doer; it implies that one stops engaging with him/her. Disdain presupposes no authority on the part of its object (Darwall 2006, 77). Unlike the experience of reactive attitudes, disdain doesn't attribute to the disdained person the capacity to act on the second-personal reason, which is connected to the authoritative demand (Darwall 2006, 77). In contrast, experiencing the reactive attitudes shows that one presupposes that the object of our reactive attitudes has the capacity to act on the second-personal reasons that he/she has violated: "When we respond reactively to someone who fails to respect a moral demand, we attribute to her the capacity to act on the distinctive kind of reason, second-personal reason, that is connected to the authoritative demand." (Darwall 2006, 77). When we experience the reactive attitudes, we are not merely demanding to be treated in a different way, but we presuppose that the individual in question is capable of recognising this demand and of recognising our authority for pressing that demand (Darwall 2006, 77).

As an illustration of his stance on disdain, Darwall brings up the particular example of two ice skating performers, Barbara Fusar Poli and Maurizio Margaglio. In one particular performance, Maurizio Margaglio slightly dropped Barbara Fusar Poli during their performance, causing her to give him at the end of the performance a very expressive stare that she addressed to him as a way of conveying her disappointment at his mistake and her asking him to account for it. Darwall contrasts this expressive stare with another kind of possible facial expression that could be adopted in similar circumstances: rolling one's eyes. Rolling one's eyes is, from Darwall's point of view, a fairly straightforward expression of disdain. It expresses having given up on the person being able to do better as it were and not taking them as having moral freedom. I will put this claim into question

below by arguing first that the ideal character of Darwall's account blocks it from having such a normative implication (see section 6.1) and second that disdain (and contempt), far from showing lack of respect, presupposes respect (see section 6.2).

C. *Undermining the moral wrongness of an action*

Finally, a more radical move would be to point out that not experiencing the reactive attitudes would undermine the very claim that a wrong has been committed. Recall that, according to Darwall, what we are morally obligated to do is, as a conceptual matter, what we are warrantedly held responsible for doing, what would warrant blame and be culpable if done without adequate excuse (Darwall 2010, 221). This could lead us to believe that the absence of reactive attitudes and the adoption instead of an attitude of disengagement, such as disdain and contempt, would undermine the claim that a wrong has been committed. However, I will argue that the fact that a wrongdoing warrants blame will not be affected by whether or not the reactive attitudes are experienced in a particular actual case. Whether or not a wrongdoing warrants blame will only be affected by whether or not it would *warrant* the experience of the reactive attitudes by *ideal* members of the moral community. I will return to this later (see section 6.1).

5. *Two objections*

5.1. *Can the SPS have normative implications?*

But before I attempt to reject the views mentioned above, let me consider straightaway a possible objection to the claim that the SPS might have *any* normative implication. To say that the SPS has a normative implication is to say that it entails a claim about what is valuable and disvaluable. However, the SPS is a meta-ethical account. Why shall I assume that one can derive substantive normative claims from the SPS? After all, it is generally assumed that one can't derive substantive normative ethics claims from meta-ethical ones.

To that I shall respond that, although the SPS is mostly concerned with meta-ethical issues, Darwall himself recognises that the arguments developed within the book might have normative implications for the content of our moral obligations:

> But what specific demands does this shared basic authority legitimate? What specifically can we reasonably expect of one another? This is obviously a normative ethical question rather than a fundamental metaethical issue of the kind on which SPS takes an explicit stand … But while SPS is largely silent on specific normative questions concerning the content of our moral obligations, I believe that its arguments have substantial normative ethical implications nonetheless. (Darwall 2010, 225–226).

If Darwall himself claims that one can take the SPS to have normative implications, then we can at least not reject as obviously wrong-headed attempts to derive normative implications from the SPS. Moreover, throughout my argument below, I will show how such normative implications seem quite naturally to follow from the SPS.

There is much more to say about the difficulties to derive a normative implication from a meta-ethical account but I won't be discussing this here. However, in one of my responses, I will point out that the SPS can't have normative implications for the value of the actual experience of disdain, because of its *ideal* character. And I will be discussing the validity of only one such ethical implication, namely, that the replacement of reactive attitudes by an attitude of disengagement, such as disdain, is necessarily morally disvaluable.

5.2. *Disdain and contempt: distinctive reactive attitudes?*

But all this presupposes that disdain is not one of the reactive attitudes. But if one challenges this claim, some of the arguments presented above might not hold. Consider the argument to the effect that disdain fails to treat the wrongdoers with respect. Let us schematically summarise one of the arguments presented above in the following way:

1. The reactive attitudes are required if wrongdoers are to be held accountable.
2. If one feels disdain towards a wrongdoer, one (normally) cannot also feel reactive attitudes towards that wrongdoer.
3. Disdain fails to hold wrongdoers accountable, and fails thus to respect wrongdoers as free moral agents.

Therefore

4. Disdain is morally disvaluable or inappropriate.

This argument could be resisted if we rejected premise (2) and showed that disdain is in fact itself a reactive attitude. Michelle Mason has argued that contempt is a reactive attitude, as it is "a reaction to another's expressing towards us or those of concern to us some quality of will" (Mason 2003, 251). But I believe that more can be said in favour of disdain than merely stating that it is also a reactive attitude. Note however that some of the arguments I will present below could be interpreted to also support the claim that disdain is a kind of reactive attitude.

I don't think, however, that the argument needs to turn on whether or not disdain and contempt are reactive attitudes (premise 2) but rather on the more important question of whether or not disdain is compatible with holding the object of disdain responsible (premise 3). If premise 3 can be held independently from the reactive attitudes account, the argument could be formulated without premise 2. The objection might thus not invalidate a similar argument provided that premiss 2 is not used.

6. *Two answers*

Let me now defend the view that disdain is not necessarily morally disvaluable or problematic. For that purpose, I will suggest two different responses to the arguments presented above to the effect that disdain is disvaluable. The first one will point out that the normative implication that disdain is disvaluable only follows from the account if we are unclear about the nature of the claims made. The claims made pertain to the disposition that ideal members would have and the account is definitely ideal. The experience of the reactive attitudes in itself doesn't ground any moral obligation, as the moral obligation is independently grounded on the disposition that hypothetical members of the moral community would have. The actual experience or absence of reactive attitudes has thus no immediate implication for moral obligation or responsibility. I point out that the ideal character of the account also undermines the validity of the argument based on self-respect.

The second answer is that, contrary to the claims made by Darwall, experiencing disdain presupposes that the victim holds the object of disdain responsible and accountable for her character traits as instantiated by her behaviour. Without the presupposition of accountability, no disdain would be experienced by the victim towards the wrong-doer. In order to

defend this position, I will introduce a distinction between normative and empirical expectation towards the wrong-doer.

6.1 *Actual vs. ideal*

The first answer would be to point out that the hypothesis presupposes that *actually experienced reactive attitudes* are what matters and what ground moral obligation. But this could be denied, and this would block Darwall's account from having the implication that experiencing disdain would undermine the very claim that a wrong has been done and that a moral obligation was violated. (As I will point out, this section will also clarify why I think that the ideal character of Darwall's account also undermines the validity of the other arguments.) Recall that I have assumed that experiencing disdain leaves no space to experience the reactive attitudes. If actually experienced reactive attitudes are required to ground moral obligation, experiencing disdain would undermine the claim that a wrong has been done and that a moral obligation has been violated. It would thus be morally problematic to experience disdain, as it would imply that no moral obligation has been violated.

Many critics have argued that the SPS should not remain vague on this issue, as taking the view that actually experienced reactive attitudes ground moral obligation would be very problematic. One such objection is articulated for instance by Jay Wallace: he argues that it is counter-intuitive to claim that I wronged you only when you raised a demand not to have your foot stepped upon, and not beforehand, that is, when your foot is being stepped upon in the first place:

> this has puzzling consequences if we take seriously the idea that it is the addressing of a claim or demand that is the source of distinctively second-personal reasons. The claim or demand that is at issue in this case is the victim's protest, which we should understand as creating a reason for you to desist, in virtue of the victim's authority to make demands of precisely this nature … this cannot be right however. Surely we want to say that you have an agent-relative reason not to step on someone's gouty toe that is (to some degree) prior to and independent of any complaint that might be issued after the toe has actually been stepped on. (Wallace 2007, 26)

In other words, Wallace argues that the SPS hinges on making moral obligation depend on the actual response of individuals in interaction with

each other (Wallace 2007, 27).[5] There are different ways in which it might be objected that the second-personal standpoint seems to make moral obligation depend on the actual response of individuals. Appropriately, there are two different responses that Darwall gives to these different ways of understanding this objection.

First, one way of understanding this objection is to see it as stating that actual individuals who have generally second-personal competence fail in this particular instance to experience the reactive attitudes. We could imagine cases in which someone is temporarily depressed or is affected by something else, to the extent that the particular wrong in question doesn't really affect her. It is particularly easy to conceive of such a case if we consider the foot stepping example discussed by Darwall. The sharp pain resulting from the foot stepping might just not elicit any reactive attitudes in a person otherwise troubled by psychological grief. It seems right to say, following Wallace, that it would be odd to conclude from the absence of reactive attitudes experienced by the victim to the claim that there was no moral obligation not to wrong the victim in the first place.

However, Darwall replies to this objection by pointing out that his account assumes that demands are addressed not only when they are articulated but whenever there is a *disposition to respond to certain norm violations with the reactive attitudes* (Darwall 2007, 65). What Darwall has in mind is thus the demand that is implicit *in the disposition to experience the reactive attitudes*. There is thus no need for the individual in question to *actually experience* the reactive attitudes for her to be wronged. It suffices that she has the *disposition to experience* the reactive attitudes in this case. The fact that the expression of this disposition is blocked by other factors doesn't undermine the existence of this disposition.

Second, the objection could take another form, and be further pressed against Darwall's account, as not every human being has even this general *disposition* to experience the reactive attitudes when wronged. The objection would then point to the counterintuitive nature of claiming that these beings can't be wronged (or held accountable) because they generally lack such a disposition. Recall that, in Darwall's view, lacking such a disposition amounts to lacking second-personal competence. This seems to imply in particular that autistic individuals and psychopaths, which Darwall take to lack second-personal competence, can't be wronged or

5. Michael Smith and Jada Strabbing voice a similar worry when they wonder whether the reason not to wrong the victim existed all along or whether it is created by the demand (Smith and Strabbing 2010, 239, fn 2).

held accountable, which seems counter-intuitive. When confronted with this version of the objection, it is not clear to me what Darwall's reply is, as he seems to resort, in different passages, to two very different kinds of response.

Darwall seems purposely to leave it open whether or not one could be accountable without having second-personal competence. He doesn't commit himself to the view that second-personal competence is necessary for moral accountability but only to the view that second-personal competence is sufficient for moral accountability. This leaves entirely open the possibility that those who lack second-personal competence could nevertheless be held accountable. However, this accountability would be grounded elsewhere, not in the second-personal standpoint.

But in many passages, Darwall invokes another kind of reply: it is not the disposition to experience the reactive attitudes, as entrenched in actual individuals, but rather the disposition that *ideal members of the moral community* would have which ground moral obligations. He says so explicitly in his reply to Wallace:

> but the moral community as I understand it is not any actual community composed of actual human beings. It is like Kant's idea of a realm of ends, a regulative ideal that we employ to make sense of our ethical thought and practice. … We might therefore understand the moral community as being prone to the reactive attitudes in a contractualist way, for example, taking it that moral demands are in force if no one could reasonably reject principles that would warrant them … (Darwall 2007, 65)

In this passage, it is made clear that what really grounds the moral demand is the disposition that hypothetical members of the moral community would have towards the reactive attitudes. But if the notion of demand implicit in the reactive attitudes we experience is crucial to determine moral obligation, it matters a great deal whether what grounds this demand is the disposition to the reactive attitudes that ideal members of the moral community would have or the experience of these reactive attitudes by actual individuals. If the demand implicit that is at stake is the one that hypothetical members of the moral community would be disposed to have, whether or not one individual experiences the reactive attitudes (or is disposed or not to experience the reactive attitudes) is not going to undermine the moral obligation. What we as actual individuals end up experiencing will not have an impact on the existence of a demand that is implicit in the disposition that hypothetical members of the moral

community would have. But if that is the case, this cannot be a ground for claiming that the experience of disdain is morally problematic, as not experiencing the reactive attitudes is irrelevant to whether or not a moral obligation has been violated.

Let me clarify the importance of this point. If, on Darwall's view, wrongness and accountability are grounded in the actual experience of reactive attitudes, then the absence of the reactive attitudes (because one feels disdain) could actually mean that the act isn't wrong, the victim hasn't been wronged, there was no obligation to act otherwise, etc. But if this grounding role of the reactive attitudes relates to an ideal community, this worry dissolves and the experience of disdain doesn't undermine any of these claims. Bringing up the hypothetical members of the moral community has thus a big impact on what kind of implications the second-personal account of morality presented in the SPS can have on disdain.

Moreover, this point also undermines the two alternative arguments briefly sketched in section 4.2. If what matters is what the members of the ideal community would be disposed to experience, the actual experience of disdain towards a wrongdoer can't have the implication that the victim lacks self-respect or that the victim disrespects the wrongdoer. Recall that what actual individuals end up experiencing will not have an impact on the disposition that hypothetical members of the moral community would have. Once again, only what ideal members of the moral community would be disposed to experience will have normative implications. So this response undermines also the validity of the argument based on self-respect and the validity of the argument based on respecting wrongdoers mentioned above, as they both claimed that the SPS would have normative implications for the value of the actual experience of disdain by individuals.

Let me add that this is a puzzling move for the general account defended in the SPS. After all, the whole point of taking a second-personal perspective is to stay grounded in actual personal interactions and what they mean to each of us. The second-personal standpoint is usually invoked to avoid taking a third-person standpoint where moral obligations might be defined in a way that seems contrary to our day-to-day intuitions. Similarly, reactive attitudes are usually invoked in order to emphasize that what matters is what kind of emotions we actually have with respect to each other. The role of the reactive attitudes in Strawson, for instance, is definitely to ground moral responsibility in our actual emotions and the intuitions that actually prevail in our interactions with each other. It is thus puzzling to take reactive attitudes to have such a crucial role in

determining our moral obligations but then qualify them as those reactive attitudes hypothetical members of the moral community would be disposed to experience. Hypothetical members of the moral community are omniscient and thereby epistemically at least occupying the third-personal standpoint. If really what is determining our moral obligation is the kind of reactive attitudes that the members of Kant's realm of ends would experience, this seems to lie far away from what is usually taken to be the prerogative of the second-person standpoint and the reactive attitudes: our actual interactions with each other.

Now this is a general issue that the account defended in the SPS needs to address, but, as I argued above, it also has implications for the much narrower question that interests me here. Not experiencing the reactive attitudes and experiencing disdain (or contempt) in their place cannot be said to be morally disvaluable, I suggest, because the actual experience of disdain cannot affect the disposition that hypothetical members of the moral community would have. The absence of reactive attitudes and the experience of disdain do not thus undermine the claim that the object of disdain has behaved in a blameworthy fashion (and similarly the experience of disdain doesn't show lack of self-respect on the part of the victim or disrespect towards the object of disdain).

Although the actual experience of disdain might not have direct normative implications, Darwall could argue that individuals ought nevertheless to *strive* to have the same attitudes that the ideal members of the moral community would have. In response to that, the only thing I will have space to argue for here is that, even if hypothetical members of the ideal moral community should feel resentment rather than disdain towards wrongdoers, it remains an open question whether actual victims should feel this way. This presumably revolves on whether disdain fails to respect the moral freedom of the wrongdoers. So let me now turn to this question.

6.2 *Disdain presupposes responsibility*

On Darwall's account, the key worry about disdain is that this attitude is incompatible with holding the object of disdain to be morally free and thus thereby lacking respect towards him/her. In this section, I will focus on putting into question that claim.

I will argue against the claim that disdain necessarily denies that the object of disdain is capable of second-personal competence. On the contrary, I will argue that, in the same way that reactive attitudes presup-

pose the second-personal competence of their target, disdain similarly presupposes such second-personal competence. Note that I will remain uncommitted here as to whether or not this indicates that disdain is also best seen as a reactive attitude. What matters for a defence of disdain and contempt as morally justified attitudes is that they are not incompatible with holding wrongdoers responsible and that they are thus not necessarily lacking respect towards wrongdoers.

Before I introduce my own arguments, I would like to briefly consider the arguments put forward by Kate Abramson in her attempt to rescue contempt from the charge that experiencing contempt towards someone is incompatible with holding this person responsible (and hence treating her with respect). Abramson introduces a distinction between localised and global forms of contempt. Whereas localised forms of contempt are restricted to a specific sphere of interpersonal relation, global forms of contempt encompass the individual whole in its evaluation. She argues that only globalised contempt might be difficult to reconcile with respecting and holding the wrongdoer responsible. In other words, as long as the contempt experienced towards someone is localised, this is compatible with holding this person accountable.

She claims that localised contempt and its behavioural tell-tale sign of avoidance makes sense *as a way of holding people accountable* for not being reliable in some particular sphere of interpersonal relations (Abramson 2009, 210). I agree with her that contempt results *from* holding individuals responsible for their actions. As I elaborate below, a negative evaluation of an individual is only possible if we take that individual to be morally free. However, I don't think that the localised vs. globalised distinction helps here. To begin with, some aspects of character in one sphere of interpersonal relation might have implications for another sphere of interpersonal relation. If someone exhibits a very selfish character while interacting with his colleagues, it doesn't necessarily imply that the individual in question will be selfish in every personal interaction he has. In practice, however, a character trait is rarely restricted to one area of interaction. But more generally, I believe that, even if an individual commits wrongdoings in various spheres of interpersonal relations, and contempt of a greater scope is warranted towards her, one could still hold her responsible for her behaviour. As I will argue below, contempt presupposes responsibility rather than exclude it.

Let me now explain why I believe that disdain doesn't entail lacking respect towards the disdained party. Recall what was said above in this

paper about the way Darwall conceives of second-personal competence. Second-personal competence is the ability to comply with norms for the right reasons, that is, because one recognises that others have authority to make claims upon us. It was also described as the ability to recognize moral reasons. But if second-personal competence is the ability to recognize moral reasons, then I don't see how disdain would not, like resentment, presuppose that its object has this ability. Disdain can only be experienced towards those who we take to be responsible, and anyone who is taken to be responsible has second-personal competence.

To show this, take the example of a schizophrenic person who, in the midst of a paranoid hallucination, kills someone. Assuming that we are convinced that he was having a hallucination, it would be indeed hard to experience resentment towards the schizophrenic. And it would be as unlikely that we would experience disdain towards him. This is so because we don't experience disdain or the reactive attitudes towards those who can't do better. The same could be said about resenting an autistic individual for not being moved enough by empathic feelings. It would be inappropriate for me to experience either the reactive attitudes or disdain towards her, because of her inability to experience empathic feelings.[6] In other words, we *can* appropriately feel disdain only towards those we hold morally responsible, and those who can be held responsible have second-personal competence. As Michelle Mason has expressed this thought eloquently: "Regarding another with contempt does not thereby objectify another person; rather, it is regarding him as beneath contempt that signals we have exiled him from the moral community with us." (Mason 2003, 263).

Another way to argue for the claim that disdain doesn't fail to show respect to the blameworthy individual and hold her accountable would be to invoke Darwall's own distinction between recognition and appraisal respect (Darwall 1977). Appraisal respect is close to the concept of esteem. It is the kind of respect I have towards individuals I admire for something they do or a character trait they exhibit. For instance, it is the kind of respect I will experience towards people I admire for their intelligence, their courage, ability to reach some difficult yoga poses or work achievements. Appraisal respect is displayed by "persons or features which are held to manifest their excellence as persons or as engaged in some specific

6. I think the case of the psychopath might be different, but I believe that the psychopath case is more controversial than these other cases. One might take the psychopath responsible and thus accountable for what he does. But we can put aside this case for now.

pursuit" (Darwall 1977, 38). Appraisal respect is not owed to all persons; it is owed to those who deserve it.

By contrast, recognition respect is the respect that one owes to all persons. It implies that "they are entitled to have other persons take seriously and weigh appropriately the fact that they are persons in deliberating about what to do." (Darwall 1977, 38). If some facts or features are appropriate objects of recognition respect, this entails that "inappropriate consideration or weighing of that fact or feature would result in behaviour that is morally wrong. To respect something is thus to regard it as requiring restrictions on the moral acceptability of actions connected with it." (Darwall 1977, 40). Moreover, this restriction would arise not incidentally but because of the feature or fact in question. In other words, it is because they are free moral agents that our actions towards them are limited by the boundaries of morality. A final distinguishing feature between the two forms of respect is that whereas recognition respect (like second-personal competence) is a range property supervening on some basic agential capacities, appraisal respect is a scalar property.

But if one uses this distinction, the charge against disdain loses its grip. Recall that whereas appraisal respect is the respect I experience towards an individual for some achievement, recognition respect, according to Darwall, is the respect I owe to all individuals. But if recognition respect is the respect I owe to all individuals, when I experience disdain towards a person, it is thus not her recognition respect which is at stake. I must also respect (in the recognition respect sense) those I feel contempt or disdain towards. When I experience disdain, I rather indicate very low *appraisal* respect. If I experience disdain towards someone who has behaved less than exemplarily towards others, it doesn't mean that there are no moral restrictions as to how to treat this person. I still owe her recognition respect. Moreover, there is nothing problematic about experiencing low *appraisal respect* towards some individuals, as it is after all the whole point of appraisal respect to regard individuals differently according to their merit. Low appraisal respect can be legitimately experienced towards those individuals who have stooped below tend to a certain standard. Moral agents who reflect on what values there are adopt such standards. If they were not to react to the crossing of such standards by endorsing the attitude of contempt, one might question their real commitment to the values they have. As Macalester Bell writes: "There is a conceptual connection between valuing and being disposed to a range of hard feelings when what one values is threatened. If you claim to value something but

you aren't disposed to feel any negative emotions when what you claim to value is in jeopardy, there is reason to doubt that you actually value what you claim to value." (Bell 2013, 161).

But, an objector might press on, this seems to make light of the fact that when I experience contempt towards someone, I stop expecting better from her. Surely ceasing to expect better from someone implies that I cease to treat her as a free moral agent? I don't think so. I believe that this only indicates to us that a distinction ought to be drawn between giving up expecting better from this person in a normative sense, thereby assuming that the person is no longer part of our moral community (i.e. the person cannot behave morally at all, has no second-personal competence), and giving up demanding and expecting better from this person in an *empirical* sense.

What is crucial here is the distinction between *normative* expectations and *empirical* expectations. Disdain might be compatible with holding certain normative expectations of the individual but not with keeping (implausible) empirical expectations that the individual will in fact do better. I have a normative expectation that individuals should, and therefore can, do better, and this would apply to every individual that I take to be, in Darwall's own terminology, a member of the community of moral equals. Every member of that community is accountable and can be held responsible for their actions. I would stop having the normative expectation that these individuals can do better only if I judge them to be outside our moral community. However, I can still legitimately lower my empirical expectations of some individuals when these expectations need to be adapted on the basis of the evidence provided by their past behaviour.

Let me take the example of Berlusconi. Let us say that I experience disdain towards Berlusconi. This disdain could result from an examination of his personal or political actions. It doesn't matter for the particular purpose here why I experience disdain towards him. This disdain, in my opinion, is justified because I believe he could behave better, that is, that he is able to do so.[7] If I believed he could not behave better, say he was rather subjected to some strong compulsion that led him to take some political decision or to behave in a certain way privately, then it would be strange for me to experience disdain towards him. By experiencing disdain, I am not saying however that I stop expecting him normatively to behave

7. Of course, I might be wrong in thinking so, but if I believed otherwise, that is, that Berlusconi was not able to behave better, I would not experience disdain towards him.

better. My normative expectation towards him is unchanged—I still take him to have full moral agency and hold him accountable for his actions. However, I don't expect him *empirically* to behave better. In fact, I would be a fool if I did. I would be lacking some fundamental common sense if I was surprised to hear that he lacked political judgment or that he behaved viciously in his private life.

In other words, I don't take disdain to be shutting off the wrong-doer from the moral community. Disdain, however, might lead us to shut the disdained person off from those we trust, love, esteem, or even interact with. But this is all right, as trust, love, and esteem are emotions that are better felt towards those who deserve them. More controversially, I believe that disdain is actually the appropriate attitude to have towards those people, like Berlusconi, whom we (empirically) expect to continue to misbehave and violate basic rules of moral conduct.

In fact, there are cases where feeling resentment towards a wrongdoer would appear naïve. To be sure, it is always permissible for the victim to experience the reactive attitudes with respect to a wrongdoing. However, experiencing resentment would be psychologically sound only if the wrongdoing comes as a surprise. But if the probability that, say, a repeat offender will commit a wrongdoing is very high, then feeling resentment towards this wrongdoer might be inappropriate on empirical grounds. First, it seems to be epistemically defective to still expect better behaviour in the face of strong counter evidence. Second, and more practically, it would just be emotionally exhausting to do so continuously. So although it is permissible to experience the reactive attitudes, there might be practical reasons in favour of substituting disdain to resentment in certain cases.

Let me give you an illustration of such a case. This case might involve the absence of resentment rather than straightforward disdain, but it is similar enough, to the extent that it presupposes an empirical expectation that the person would not do better. Some time ago, Obama was reported by the media to have said that he "couldn't bother getting angry" about Netanyahu's agreeing to new settlements on the West Bank. Of course, one could read this "couldn't bother getting angry" itself as an expression of resentment and anger. But if we take the exclamation to be a more honest reflection of Obama's actual state of mind, then this would exactly illustrate the point made above. The empirical expectation that Netanyahu would make such a decision was very high, and to become angry at the news would have indicated naivety. "Bothering getting angry" in this case would have indicated that Obama had failed to take into account evidence

from the past to draw conclusion about the likelihood of a certain behaviour in the future. Obama's empirical expectations have adjusted to what is likely for Netanyahu to do and achieve, but that doesn't mean that his normative expectations have adjusted. Obama still thinks that Netanyahu *should* have prevented the building of new settlements on the West Bank, and this remains his normative expectations towards Netanyahu.

7. *Conclusion*

The second-personal account of morality developed by Darwall in *The Second-Person Standpoint* seems to imply that the attitudes of disdain and contempt are morally problematic and disvaluable. Darwall indeed explicitly argues that disdain is problematic to the extent that it doesn't treat its object as morally free, that is, as able to do better. This view seems to prima facie follow from the second-personal account developed in the SPS.

A fundamental claim of the account is that the reactive attitudes issue a demand to be treated with respect. But if that is the case, the withdrawal of the traditional reactive attitudes of resentment and anger and their replacement by an attitude of disregard (whether contempt or disdain) can appear to give up on making that demand. Giving up on the demand to be treated with respect has been taken to be problematic for various reasons: it might show lack of self-respect (Murphy 1982), it might fail to treat the wrongdoer as morally free and belonging to our community of moral equals (Darwall 2006, Hill 2000), it might fail to show the respect we owe to every individual (Kant 1991), or it might undermine the very claim that a moral obligation has been violated (Wallace 2007).

In this paper, I have suggested two possible ways to reject the claim that disdain is necessarily disvaluable. First, I have argued that one cannot derive from the account defended in the SPS these normative implications about disdain, as the account grounds moral obligation and moral responsibility on the dispositions of hypothetical members of the moral community. In Darwall's view, the second-person standpoint is constituted by the perspective of hypothetical members of the moral community, and moral obligation is constituted by what would warrant their reactive attitudes. To that extent, the actual attitudes of individuals are irrelevant. In other words, the SPS is pitched at the wrong level to have the direct normative implications about disdain that have been thought to follow from it.

Second, I have argued that, even if we granted that the SPS could have such implications, and contrary to what is argued in the SPS, disdain doesn't fail to presuppose the ability of its object to change, or fail to treat its object with respect. This becomes clear when one considers the matter in terms of Darwall's distinction between recognition and appraisal respect as well as in light of the distinction I have introduced between empirical expectations and normative expectations. Disdain can thus be seen as having two main virtues. First, it is an attitude that is fact-sensitive, it allows agents to adjust their empirical expectations to others' wrongdoings. Second, it fulfils the important function of restricting the number of agents towards which we experience the emotionally demanding reactive attitudes. By lowering our empirical expectation towards those we experience disdain towards, we save ourselves much emotional energy.

Bibliography

Abramson, Kate 2009: "A Sentimentalist Defence of Contempt, Shame, and Disdain". In: Peter Goldie (ed.), *The Oxford Handbook of Philosophy of Emotion*. Oxford: Oxford University Press, 189–213.

Bell, Macalester 2005: "A Woman's Scorn: Towards a Feminist Defense of Contempt as a Moral Emotion". *Hypatia* 20 (4), 80–93.

— 2013: *Hard Feelings, the Moral Psychology of Contempt*. Oxford: Oxford University Press.

Ben Ze'ev, Aaron 2001: *The Subtlety of Emotions*. London: MIT.

Capes, Justin 2012: "Blameworthiness without Wrongdoing". *Pacific Philosophical Quarterly* 93(3), 417–437.

Darwall, Stephen L. 1977: "Two Kinds of Respect". *Ethics* 88 (1), 36–49.

— 2006: *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.

— 2007: "Reply to Korsgaard, Wallace and Watson". *Ethics* 118(1), 52–69.

— 2010: "Precis: The Second-Person Standpoint". *Philosophy and Phenomenological Research* 81 (1), 216–228.

Ekstrom, Laura W. 2000: *Free Will: A Philosophical Study.* Boulder, CO: Westview Press.

Fischer, John Martin & Ravizza, Mark 1993: *Perspectives on Moral Responsibility*. Ithaca, NY: Cornell University Press.

Hill, Thomas 2000: *Respect, Pluralism and Justice: Kantian Perspectives*. New York: Oxford University Press.

Hurley, Elisa A. & Macnamara, Coleen 2011: "Beyond Belief: Towards a Theory of the Reactive Attitudes". *Philosophical Papers* 39 (3), 373–399.
Kant, Immanuel 1991: *The Metaphysics of Morals*. New York: Cambridge University Press. Translation by Mary Gregor.
Mason, Michelle 2003: "Contempt as a Moral Attitude". *Ethics* 113 (2), 234–272.
Murphy, Jeffrie G. 1982: "Forgiveness and Resentment". *Midwest Studies in Philosophy* 7, 503–516.
Nelkin, Dana 2012: *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.
Rawls, John 1999: *A Theory of Justice* (rev. ed.). Cambridge, MA: Harvard University Press.
Smith, Michael & Strabbing, Jada 2010: "Moral Obligation, Accountability, and Second-Personal Reasons". *Philosophy and Phenomenological Research* 81 (1), 237–245.
Strawson, Peter F. 1968: *Studies in the Philosophy of Thought and Action*. London: Oxford University Press.
— 1974: *Freedom and Resentment and other Essays*. London: Methuen.
Wallace, R. Jay 1994: *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
— 2007: "Reasons, Relations, and Commands: Reflections on Darwall". *Ethics* 118(1), 24–36.
Watson, Gary 2005: "Responsibility and the Limits of Evil". In: John Martin Fischer (ed.), *Free Will Concepts and Challenge*. London: Taylor & Francis, 106–135.

Grazer Philosophische Studien
90 (2014), 105–129.

SELF-CONSTITUTION AND OTHER-CONSTITUTION: THE NON-OPTIONALITY OF THE SECOND-PERSON STANDPOINT

Christoph HANISCH
University of Vienna

Summary

A second-person standpoint theory of practical normativity can be defended against the criticism that it issues merely conditional requirements. This (Neo-Kantian) criticism claims that taking up the second-person standpoint is something that an agent can avoid, while this is not true with respect to her first-person, deliberative, perspective. I employ a social extension of recent work on self-constituting action in order to show that this is not the case. Self-constituting action (in the presence of other agents) depends for its possibility on shared norms and practices of (at least) non-interference, which in turn renders presuming "second-personal competence" non-optional.

Introduction

Stephen Darwall has presented an inspiring account of practical normative requirements, moral obligation in particular[1] (Darwall 2006). According to his second-person standpoint theory, persons take up a unique perspective whenever they engage in making and acknowledging interpersonal demands and claims. When this perspective is taken up, individuals are thereby subject to a particular category of normative requirements and,

1. It will be one of the features of my account that it is not concerned with the distinctively moral sphere, understood as a subset of all normative requirements. The account of the non-optionality of the second-person standpoint, presented in the text, therefore differs from Darwall's (and Korsgaard's first-personal alternative) in that it does not claim to be committed to vindicating all the substantial moral conclusions that *The Second-Person Standpoint* strives to establish. I am concerned with a more abstract and general account of second-personal normativity that is well compatible with morally problematic actions, attitudes, etc. However, nothing in what follows rules out that the defense of the second-person standpoint's non-optionality in terms of normative requirements *simpliciter* can be developed into a richer and more ambitious direction.

equally important, normative presuppositions that agents must presume *vis-à-vis* one another in order for the aforementioned demands and claims to be intelligibly made and acknowledged. When engaging in such "second-personal address" the addresser must, for example, presuppose in others (the addressees) certain cognitive capacities and normative competencies, such as that they can subject their conduct to practical requirements and that they can guide their choices and actions accordingly.

Critics have argued that this picture of the sources of normative requirements introduces a problematic contingency and lets these peculiar normative requirements rest on a set of merely conditional commitments. We always seem to be in a position to ask, "But why am I under a categorical requirement to take up the second-person standpoint in the first place?" In other words, the second-person standpoint, and hence all the normative baggage that it presupposes and commits one to, appears merely optional. Moreover, at the same time, taking up the first-person standpoint, the standpoint from which every deliberating and acting agent chooses her actions, is not optional in this way. According to this asymmetrical picture, engaging in second-personal address merely appears to be an optional "add-on" that commits one to the elaborate structure of normative requirements that Darwall develops only in so far as an agent chooses to take this perspective seriously.

In this paper I defend the second-person standpoint as non-optional and do so in the framework of a socio-ontological investigation concerning the prerequisites of individual agency and identity. In reply to, amongst others, Christine Korsgaard, who argues for the asymmetry between the non-optionality of the first- and the second-person standpoints, the argument further develops some of Korsgaard's own central ideas concerning self-constitution and practical identity. It is argued that the first- and the second-person standpoints are "normatively symmetrical," in the sense of conditioning one another. In section two, interpersonally shared normative practices of non-interference are identified as necessary conditions of individual self-constitution. The non-optionality of presuming second-personal competence in others is a corollary of the claim that non-interference practices are indispensable for action. In sections three and four, two objections to this argument are considered respectively. In the course of presenting the reply to the first objection a revised notion of "second-personal address" (one that accommodates agents who seemingly don't engage in it at all) is presented. The reply to the second, deeper cutting, objection argues against the claim that the acknowledgment of other

human beings *qua* second-personally competent agents is merely optional from the point of view of what is required for successful individual self-constitution. It is ultimately my own practical agency's normative features that rule out reducing "second persons" to mere forces of nature.

I. *The problem of optionality*

One criticism that has been leveled against Darwall's conception of practical normativity is that its central element, i.e., the second-person standpoint, is a stance that agents need not necessarily take up. According to Darwall, the second-person standpoint is "the perspective you and I take up *when* we make and acknowledge claims on one another's conduct and will" (Darwall 2006, 3; my emphasis). Darwall himself recognizes that his conception of practical requirements, that generates its normative force from the presuppositions of the justificatory activities that take place among interacting parties, faces a problem very similar to the one that Kant has to deal with in *Groundwork III*: while the first part of Darwall's argument in *The Second-Person Standpoint* presents a thoroughgoing analysis of mutually connected and reinforcing normative concepts that are all ultimately related to the second-person standpoint, he admits that "(e)ven if taking up the second-person stance commits us to equal dignity and autonomy, that is consistent with that standpoint and its associated commitments being no more than rationally optional, or worse, illusory" (Darwall 2006, 277). Darwall relies on Fichte in order to develop a response to this challenge. Taking up the second-person standpoint, so Darwall,

> gives us a perspective on our own agency that enables us to appreciate a fundamental difference between theoretical and practical reason and so improves our grasp of reasons for acting. Were someone somehow to avoid taking it up, consequently, she would fail to appreciate what we, who have taken it up, can validate as reasons from a more comprehensive view that includes it (Darwall 2006, 277).

Amongst others, Christine Korsgaard has expressed worries concerning Darwall's proposed solution to the optionality-problem.[2] She continues to press the point, summarized above by Darwall, by emphasizing the

2. See Korsgaard 2007, Pauer-Studer 2010, and Schapiro 2010.

one aspect of our nature as practical reasoners that her Kantian account regards as foundational: "(T)he second-person standpoint does not seem to be unavoidable, the way the standpoint of first-personal deliberation is" (Korsgaard 2007, 22). It is the crucial feature of Korsgaard's account of normativity that it locates the origin of normative requirements squarely within the viewpoint of the deliberating agent (and within the constitutive features of that viewpoint) (Korsgaard 1996, Lectures III and IV). In spelling out her worries concerning the optionality of second-personal features of our practical identities, Korsgaard agrees with Darwall on two counts: the psychological difficulties that come with the attempt to avoid the second-person standpoint are insufficient to address the philosophically relevant questions (Darwall 2006, 138ff. and 277f.). Moreover, Korsgaard grants that "the person who declines to take up the second-person standpoint fails to know something intimate and important about his own agency" (Korsgaard 2007, 22).[3] However, and this remains the main point of contention between Darwall's account and first-person theorists, an agent can always ask herself why she should take up the second-person standpoint with all its presuppositions and normative commitments in the first place. In other words, while the first-person deliberative standpoint seems to be self-sufficient and freestanding (*pure* first-personal practical reflection and deliberation appears to be a coherent possibility), second-personal engagement and address merely count as "add-ons" that agents need to have some positive reason to take up and to regard as normatively potent.

The last sentence does not do full justice to the Neo-Kantian strategy of dealing with Darwall's problem though. Both Korsgaard and Tamar Schapiro broadly adopt the same strategy to "rescue" the second-person standpoint from the criticism that it fails to provide unconditional and categorical requirements. They argue that, in a sense, this standpoint *is* non-optional after all, namely in the form of "the voice of the second person within" (Korsgaard 2007, 23) and because of the second-personal character of our desires, exposure to which we experience as a form of "self-division" (Schapiro 2010, 235). The common point of these two

3. This weakening (at least I consider it as such in this paper) of his main point is also expressed in the long quotation above in which Darwall describes agents who take up the second-person standpoint as (merely) gradually "improving" their grasp of reasons that the alternative, non-second-personal agent, seems to be capable of grasping too (only to a lesser degree). Such an agent, and she seems to be considered a proper, well-constituted, agent, merely fails to validate "reasons from a *more* comprehensive view."

strategies is that practical deliberation, conceived from the first-personal perspective, takes on the form of an inner encounter and "dialogue" in the course of which an agent constitutes herself into someone who has a coherent and consistent practical identity. This inner conversation, as opposed to making and responding to demands of actual other agents, inescapably shows an agent that she is, first and foremost, answerable to herself (not necessarily to others). Here is Korsgaard:

> [T]he reflective structure of self-consciousness inevitably places us in a relation of authority over ourselves and that we are as a consequence also accountable to ourselves. … [E]very rational agent stands in what Darwall would call a second-personal relation to herself—she has a second-personal voice within (Korsgaard 2007, 11).

Darwall's responses to these proposed solutions to the optionality problem seem to concede to his objectors quite a lot—but not straightforwardly so.[4] On the one hand, Darwall continues to defend a strong, Fichtian, version of his central claim concerning the non-optionality of the second-person standpoint when he states that "we unavoidably take up a second-person perspective when we respond to address (someone else's or our own) as a Fichtean *Aufforderung*" (Darwall 2007, 59f.).[5] It is in passages like this one that I see Darwall's view getting close to suggesting that the actual presence of another practical reasoner is not merely an (empirically) contingent feature of one's normative condition. When actually responding to and interacting with others, a genuine sphere of categorical and universal normative (and moral) requirements presents itself. As will be spelled out more fully below, it is a sphere that depends precisely on the other person(s) being a source of normative requirements that is *independent* of the deliberating agent, her will, and her purely first-person viewpoint.

On the other hand, however, immediately following passages like the one just mentioned (as well as indicated in the parenthetical "or our own"

4. See Darwall 2007, 59-60 and Darwall 2010, 255ff.

5. That Darwall here allows a Fichtean *Aufforderung* to be both "someone else's *or* our own" address complicates the reading I suggest in the text because granting the second disjunct seems to render his account indistinguishable from Korsgaard's "second-person within" approach. In the text I will present considerations for why "our own address" cannot be a sufficient substitute for encountering the independence of other agents and their volitional capacities. It is by means of defending the stronger reading that we present a way of introducing a sharper contrast between Darwall's and first-person standpoint theories.

therein), Darwall softens the claim of what it means to regard the second-person standpoint as non-optional. In a passage in which he situates his claim in metaethical controversies between realism and constructivism, Darwall states that "(e)ven if it is adopting the second-person standpoint that commits us to accepting the existence of second-personal reasons, that does not mean that second-personal reasons themselves depend upon our actually taking up a second-person point of view" (Darwall 2007, 60). Moreover, in other passages (and in personal correspondence) Darwall insists that the introduction of the second-person framework into practical philosophy is not meant to identify the actual confrontation of a multitude of individuals with one another as somehow necessary for the possibility of normativity; rather, Darwall's concern is to create a "logical and conceptual space" (for a particular kind of normative items) that he considers neglected in his contemporaries' theories. The question is whether or not this strategy makes Darwall's version of a second-person standpoint theory collapse into the Kantian, first-person, ones that he criticizes as deficient.

Just recall Korsgaard's "second-person within," according to which second-personal reasons are perfectly well constructible from within one's own deliberative, first-personal, standpoint. If this is what Darwall means by establishing the second-person standpoint as a distinct logical and conceptual sphere of practical reasoning, then the problem is that it seems to render the second-person standpoint redundant as a freestanding and autarkic source of normative requirements (Pauer-Studer 2010, 299f.). Of course, Kantians will readily grant, hypothetically taking others and their interests into consideration when deliberating about what to do is a constitutive feature of willing as the Kantian imperatives indicate (just think of the Kingdom of Ends Formula of the Categorical Imperative). However, actually engaging in second-personal address in order to do so is not strictly necessary for the phenomenon of normativity to emerge. And this is the same conclusion that Darwall establishes when he defends the second-person standpoint as a logical space only (as opposed to an interpersonal practice). This space is perfectly well accessible from within the perspective of a Kantian practical reasoner engaged in a "dialogue" with a hypothetical you. In other words, the second-person standpoint (its presuppositions and its requirements) collapses into and can be fully reduced to the first-person standpoint of a deliberating agent that Kantians identify as non-optional. Practical reflection about "The Kingdom of Ends" can be conducted solipsistically and the construction and/or discovery of

second-personal reasons might well be possible without ever conceiving of others as *independent* willers.

My claim is that Darwall cannot rest content with this result about the self-sufficiency of the first-person standpoint as Kantians seem to be able to defend it. In important passages he notes that his theory "rules out … first-personal thought that lacks an addressing, second-personal aspect. Thus although second-personal address is always also first-personal, it is never merely first-personal" (Darwall 2006, 10). Everything hinges on the word "merely" here. It is my impression that the above Neo-Kantian criticisms of *The Second-Person Standpoint* capitalize on Darwall's reluctance to commit himself to a stronger reading of second-personal address and, hence, of everything that is implicated in the latter. Schapiro, for example, sums up that Darwall "rejects the strategy of arguing that the second-personal standpoint is forced upon us by our nature as reflective agents" (Schapiro 2010, 229). This diagnosis seems correct, and it is Darwall's modesty concerning the forceful nature of having to take up the second-person stance, expressed in Schapiro's summary, that creates the problem concerning the unconditional bindingness of normative requirements—not merely moral ones—currently discussed.

In the remainder of this paper, I will argue that Darwall sells himself and his theory short when he concedes that second-personal reasons well "exist" without anybody taking up the second-person standpoint and when he concedes that a Fichtian *Aufforderung* can be formulated (and be dealt with) satisfyingly entirely "from within." As we shall see, the non-optionality of the second-person standpoint can be established by looking at recent work by the philosopher who has presented the optionality-worry most forcefully, namely Christine Korsgaard.

II. *Self-Constitution action in the presence of others*

The following defense of the non-optionality of the second-person standpoint then will have to differ from Darwall's own version of this view in so far as it regards a *second*-personal theory of normativity as requiring more than just making room for a specific logical and conceptual space. Both a synthesis and a social extension of the crucial elements of Christine Korsgaard's Neo-Kantian accounts of normativity and agency are the defense's foundation. It is a *synthesis* of Korsgaard's two central ideas because it combines her conception of an agent's "practical identity" (Korsgaard

1996, Lecture III), on the one hand, with her recent claim that "action is self-constitution" (Korsgaard 2008 and Korsgaard 2009) on the other—a combination that Korsgaard does not explicitly consider.[6] It is, at the same time, a *social extension* of these ideas because, as will be argued in this section, there is something irreducibly *inter*personal involved in the inescapable activity of constituting oneself into an individual agent (with a particular practical identity) by means of practical deliberation about and performance of actions.

According to the Korsgaardian framework, human beings are equipped with a unique form of reflective self-consciousness that, at the same time, is the source of a set of practical problems that distinguishes us from all other animals. The deliberative standpoint, the first-personal perspective that we (must) take up when we deliberate and decide what actions to perform/not perform, is not necessarily and automatically "unified," as is the case with animals whose instincts take care of that task. In the case of animals, instincts "guide" behavior and determine the practical options that a specific organism has in response to the needs and urges that it encounters. The structure of human self-consciousness, on the other hand, is marked by a gap between an occurring desire and the choice to take it up as a reason for action. The process of deliberation and choice, conceived from within the agent's perspective, is not determined by instincts, and this "practical freedom" compels us to employ practical principles in the aforementioned tasks. Practical principles are two things at once: they guide and structure our practical deliberation and action, on the one hand, and, in so doing, these sets of principles are the building blocks of our practical identities *qua* unique individual agents, on the other. In the abstract and austere rendering of "practical identity," that I use for my purposes, this technical notion simply refers to an agent's normative self-conception, a conception of being an individual who is subject to certain norms and principles (not necessarily morally justified ones) that circumscribe what actions are deemed permissible, obligatory, etc. by her.[7]

---

6. David Enoch observes that "the relation between the views expressed in the Locke Lectures [now published as *Self-Constitution*; C.H.] and in *Sources* is not entirely clear to me" (Enoch 2006, 171). Indeed, Korsgaard clarifies the relationship between her two projects (*The Sources of Normativity* and *Self-Constitution*) only very briefly. It remains an open question whether Korsgaard would agree with the synthesis aspect of my rendering of her views. On the other hand, I am quite confident that she would not agree with the social extension of her view that I suggest. See Korsgaard 2011, 392ff.

7. A recent definition by Korsgaard reads: "Our conceptions of our practical identity govern our choice of actions," and as a possessor of a particular practical identity I "find it worthwhile

In Korsgaard's recent argument concerning the constitution of agency, it is the two Kantian imperatives that figure as a kind of "meta norm" and higher order principles in this process that she now sums up in the slogan "action is self-constitution." The dispersed elements of our volitional capacities (the cognitive aspects as well as our desires) need to be unified and rendered into a coherent and consistent first-personal viewpoint and this task is accomplished by means of deliberation and action in accordance with the aforementioned practical principles that guide these processes.[8] In their form as mere principles of self-constitution (as opposed to substantive norms of morality), the two imperatives summarize the most foundational normative requirements that every action must satisfy in order to successfully fulfill its task, which is to constitute the human who performs it into a well-unified agent with a coherent and intelligible practical identity.

The categorical imperative, on the one hand, states that an agent must adopt some principle of choice lest she turn into, what Korsgaard calls, a "particularistic willer," that is, someone whose "actions" have no practical implications whatsoever for her present and future choices (Korsgaard 2009, 72–77). In other words, the categorical imperative, so understood, merely consists in the requirement to act for a reason because doing so is the only way to constitute oneself into a temporarily extended and stable agent as opposed to a heap of unstructured and un*principled* desires and urges that randomly result in behavior. In a similarly spartan manner, the hypothetical imperative sums up the fundamental idea of instrumental rationality, namely that an action that successfully constitutes an agent satisfies the requirement of taking up the means necessary to attain a proposed action's end.[9] (Korsgaard 2009, 68–71) Ultimately, therefore, it is the inescapability and non-optionality of the "human plight" (i.e., the task

---

to do certain acts for the sake of certain ends, and impossible, even unthinkable, to do others." And a little bit later: "One might think of a particular practical identity, …, as a set of principles, the dos and don'ts of being a teacher or a citizen, say" (Korsgaard 2009, 20 and 21).

8. That a practical identity and unified agency are the condition as well as the result of the process of self-constitution has been discussed intensively under the heading of "the paradox of self-constitution" that Korsgaard acknowledges herself and tries to argue away. See Korsgaard 2009, 41–44 and Hanisch 2015 [forthcoming].

9. See Korsgaard 2009, 68–71. This unduly sketchy picture of the Korsgaardian view is all we are going to assume in the following account. Keep in mind that we are arguing on the abstract level of mere normative requirements that are, first-personally, implicated in the activity of self-constituting action and identity formation. Korsgaard's extension of her basic framework into genuine moral territory (her arguments from the public nature of reasons and her analogy with Plato's well-constituted soul respectively) is not at all presumed here. I will come back to this issue of restricting my account to the "non-moral," though normative, sphere below.

of having to constitute oneself into an agent) that grounds the normative force of those practical principles that structure our practical identities as agents and, *a fortiori*, of the two Kantian imperatives that normatively constrain all self-constituting action.

In order to see how this Korsgaard-inspired account of practical deliberation and action leads to the inescapability of taking up the second-person standpoint, we need to examine the objects of choice that are involved in the activity of self- and identity-constitution more closely. Actions, understood as the vehicle of this activity, are deliberated about and chosen on the basis of the principles that determine under which normative description we see ourselves. These principles are employed and implicated in the actions that we perform. Put a bit bluntly, "we are what we do" (and "we do what we do on the principled basis of who we are"), and the claim is that whenever agents formulate the "maxim" of a proposed action (the subjective principle that describes it as something that can figure as the object of practical choice), they check its normative status ("Is this action something I can see my-self doing?") by means of mobilizing *their* practical identities, compiled of their practical principles of choice and action. It is in re-affirming those practical principles that one holds as well as in endorsing new ones (in the context of and on the basis of the former though) that one's agency and particular identity are maintained.

Now, the main claim submitted here is that the non-optionality of self-constitution provides a stringent basis for, firstly, taking up the second-person standpoint and, secondly, for conceiving of oneself, first-personally, as being subject to certain interpersonally maintained norms. In order to see this, we must keep in mind that construing an action's maxim always involves putting together an act and an end. Especially the latter task, setting ends, depends for its intelligible confrontation on certain external regularities that constitute the stability and predictability necessary to *set* and *will* an end (as opposed to merely desiring it). These regularities come in two varieties. On the one hand, determining oneself as someone who can bring about a certain state of affairs depends on the natural world behaving in a non-random way. (In order to act successfully, the ground below my feet must not shift erratically.) On the other hand, and this is the feature of the external world most relevant for our argument, surrounding agents pose a potential source of interference that endangers every agents' self-constituting activities, especially her attempt to conceive of certain ends as something that can be brought about by acting in a specific way.

Admittedly, the scenario envisioned here is an extreme one. It rests on a hyper-chaotic environment in which potential agents constantly thwart their action-attempts and, so to speak, mutually annihilate their agency by means of undermining their attempts to set ends (and to act). Still, I claim, this thought experiment draws out an intuition on the conceptual level concerning the relationship between the *complete* absence of any shared interpersonal norms, on the one hand, and the prospects for regarding *oneself* as engaging in self-constituting action, on the other.

In so far as successful action requires for its possibility a *minimum* degree of principle-based non-interference amongst agents, we can say (in so far as we stay in Korsgaard's framework) that successful self-constitution does so too. Even in the case of trivial actions that seem to be completely independent from complex interpersonal practices (like walking down the street to get to a friend's place) the process of practically deliberating about them and performing them involves shared practices of, at least, non-interference. Setting oneself the end "getting to my friend's place," and combining it with the act of walking, is not merely dependent for its intelligible construction on being structured in accordance with the already mentioned Kantian requirements of practical reasoning. Moreover, and in addition to nature acting and responding predictably, the behavior of all other agents must be *structured* on the basis of some mutually acknowledged principles and norms. That I intend to choose and endorse the action of walking down the street in order to visit my friend necessarily incorporates as one of its presuppositions that others won't interfere with me doing so in an arbitrary and chaotic manner. In the same way in which I can count on the physical make-up of the environment to remain predictable and stable, my practical deliberation presupposes that certain interpersonally shared principles of non-interference are held constant.

The gist of this argument is that we can conclude from the aforementioned relationship between shared norms and the possibility of action that the successful self-constitution of agents into beings with a coherent and stable first-personal stance too depends on there being some such shared practices in place. The practical principles that are in play when I deliberate about even seemingly trivial actions always incorporate some minimally interpersonal norms. Elsewhere I argue that this relationship between coherently taking up a first-person standpoint and the presence of external norms allows us to declare every successfully constituted practical identity a "public (practical) identity" because, in the presence of other

agents, a practical identity always also incorporates "public" principles, in the very minimal sense of principles that we, together, acknowledge as circumscribing spheres of non-interfered-with action. I cannot pursue these issues here (Hanisch 2013).

In the remainder of this paper I want to focus on another presupposition that is implicated when agents constrain the option to interfere with one another on the basis of such public principles. This presupposition is the one that gets us to what I take to be one of Darwall's most central and fundamental elements of his system, namely his account of second-personal competence. Darwall says that

> those we address can guide themselves by a reciprocal recognition of the second-personal reasons we address and our authority to address them, that they can take a second-personal perspective on themselves and act on reasons they accept from that point of view (by making the relevant demands of themselves) (Darwall 2006, 75).

When the above argument insists, *pace* Korsgaard, that assuming a second-person standpoint merely "within one's first-personal deliberative standpoint" is insufficient to confront the task of self-constitution successfully, the presumption of second-personal competence in others is the crucial linchpin underlying this paper's main thesis, viz. that self-constitution is other-constitution and the other way around. Leaving aside issues of substantive morality (certainly implicated in Darwall's definition of second-personal competence), regarding all other agents as the kind of beings who *can* subject their potential interference with others' actions on the basis of practical principles is an inescapable rational commitment, implicated whenever agents constitute themselves successfully. After all, in order for the practical norms and principles that enable oneself to engage in self-constituting action to be shared *normative* constituents of one's identity, all other agents must be presumed to be capable of incorporating (and capable of rejecting—see the next paragraph) these public principles as elements into their practical standpoints and identities. Constituting others as addressees of such norms is non-optionally implicated in one's own practical deliberation and action in their presence *in so far as* it results in successful instances of action.

When we reflect on a remark on Darwall's argument by Sam Fleischacker, we can again see why Korsgaard's suggestion concerning the sufficiency of a purely internal second-personal address is insufficient for our (and her own) theoretical purposes. Fleischacker says that "taking up the

second-person standpoint requires us to open ourselves to the possibility that a 'you' might tell us something we would never have come up with on our own" (Fleischacker 2009, 121). And, while agreeing with Darwall that second-personal address presupposes addresser and addressee conceiving of themselves as passing a certain threshold of equal "addressability,"[10] Fleischacker emphasizes that an equally important "presupposition of the second-person standpoint is that I see you as *different* from me. Otherwise I will simply project myself into you. … I will address you as if you were another 'I', not an independent being" (Fleischacker 2009, 122).

These remarks draw out an important feature of a second-person theory of normativity. It is exactly the independence (from my own person and will) that only other agents can manifest and that lifts the normativity of a practical identity's principles onto a level that a solipsistically conceived first-person standpoint on its own cannot achieve, and necessarily so. The external normative limits that we set one another, and that become constitutive features of our practical identities, therefore depend on two presuppositions that seem to pull into opposite directions: On the one hand, coherent first-personal deliberation under norms depends on conceiving of others as similar enough to oneself with respect to the capacity of together subjecting ourselves to shared practices. On the other hand, however, the shared normativity of the interpersonal prerequisites of self-constitution requires us to conceive of all others as independent beings whose contribution to maintaining practices (of non-interference) is ultimately up to them, not to me. Korsgaard's and Schapiro's phenomenology of "normativity within" neglects this non-substitutable feature of second-personal address and competence.

In summary, since some shared principle-based practices of non-interference are a necessary precondition for successful self-constitution amongst agents who acknowledge each other as such[11], the presuppositions for *sharing* the practice-defining norms too are necessarily presumed in this way. The non-interference principles that form the necessary ("external") component of *every* individual deliberative standpoint must be relevantly

10. Another issue that cannot be taken up in this essay is Darwall's notion of equality that he suggests to be implicated in the presuppositions of second-personal address. While I agree that addressers and addressees must regard one another as having "equally" passed a certain threshold of capacities and competencies when they address demands to one another, I am less convinced that anything stronger and more substantive can be developed on this austere basis alone. This, again, mirrors my reservations concerning Korsgaard's parallel project of deriving substantive Kantian morality from what is constitutive of (mere) rational agency, first-personally conceived.

11. See section IV for the relevance of this rider.

similar across distinct individuals—otherwise they would fail to provide the predictability, stability, etc. of our standing *vis-à-vis* one another that, according to the above argument from interpersonal chaos, is required for any instance of end-setting, acting, and, hence, action. And for these principles to be practical norms they must figure in the practical activities of beings who mutually presume the second-personal capacities required to deliberately subject their choices and actions to such principles. In short, it is exactly because we need to live as separate individuals that we must commit ourselves to some minimally shared practices and institutions. More importantly for our purposes, in order to accomplish the latter feat we must presume in distinct others the same that we presume in ourselves when we conceive of us as being subject to practical norms, namely those reasoning- and volitional-capacities that allow to put our standing *vis-à-vis* one another on a norm-guided basis.

III. *Leave me alone! Second-Person address amongst hermits*

Many a reader will regard this paper's main thesis as overblown. It seems counterintuitive to award the two standpoints, first- and second-person, equal importance by identifying them as normatively symmetrical and as conditioning one another. Clearly, the objector insists, we have to be well-unified first-personal deliberators *before* (conceptually speaking) the second-person standpoint can be taken up when engaging in second-personal address with others. This line of reasoning also seems to underlie Korsgaard's formulation of her critique of Darwall, outlined in section one, when she states that we can always ask for "a reason to take up the second-person standpoint and its presuppositions" (Korsgaard 2007, 22). (With respect to the first-personal stance, on the other hand, this question cannot intelligibly be asked since this would amount to regarding the task of self-constitution as optional.) Furthermore, when we keep in mind that, according to Korsgaard's theory of practical reasons, successfully constituted agency and practical identity are a necessary condition for having any such reasons at all, it might appear to make a lot of sense to agree with her prioritization of the first-person standpoint.[12] The sought-after reason for

12. Korsgaard writes: "Such identities are the sources of our reasons. … [t]hey govern choice … [t]hey are standing sources of incentives, as well as principles in terms of which we accept and reject proposed actions" (Korsgaard 2009, 21 and 22). And a little bit later, Korsgaard connects her account of reasons with the inescapable task of self-constitution, i.e., the human

taking up the second-person standpoint too then is the result of practical deliberation, conducted from within one's first-person perspective which is, in principle, sustainable without engaging any other agents.

Darwall is right when he rejects this entire question, i.e., the one asking for a positive reason to take up the second-person standpoint. He says,

> [t]he way we get into the second-person standpoint is not by seeing a non-second-personal reason for doing so and then taking it up. These would not give us reasons on which we could genuinely take it up anyway. They would be 'reasons of the wrong kind' (Darwall 2007, 59).

The main argument presented in this paper not only agrees that seeing some "non-second-personal reason" is inadequate grounds for getting us into the second-personal realm. Rather, it suggests that asking for any kind of reason for taking up the second-person standpoint is as implausible as asking for such a reason with respect to assuming the first-person, deliberative, standpoint. Since both standpoints are non-optional features of successfully constituted agency (and the one standpoint depends for its intelligibility on the presence of the other), there cannot be a pre-agential reason *to* take up either of the two. Korsgaard agrees with this claim only as long as we are concerned with taking up the first-person standpoint (just recall her presentation of the human predicament of unifying one's volitional capacities into one, principle-structured, practical identity as a "plight" that humans must confront successfully in order to have any practical reasons for action at all).

There is a scenario that seems to support Korsgaard's argument, according to which positive reasons for taking up the second-person standpoint are necessary (and possible) in a way that cannot be made sense of with regard to first-personal, self-referential, address and its presuppositions. Neo-Kantians might point towards a scenario in which (successfully constituted) agency is present, without that implying the presence of any actual second-personal address and, hence, without shared norms or practices. In so far as the latter are absent in such a scenario, while at the same time the task of identity-constitution is confronted successfully, we seem to be able to conclude that second-personal address is not necessary for

---

plight: "We must act, and we need reasons in order to act. And unless there are *some* principles with which we identify we will have no reasons to act. Every human being must make himself into someone in particular, in order to have reasons to act and to live. Carving out a personal identity for which we are responsible is one of the inescapable tasks of human life" (Korsgaard 2009, 23f.).

self-constitution and identity-formation (even in the presence of others), and we are back with its optional character and can ask for an additional reason to engage in it. The way hermits conduct their lives seems to provide a straightforward counterexample to the claim that shared practices and norms have to figure in an agent's practical deliberation and action. Furthermore, then, such a scenario suggests that an acknowledgment of the presuppositions of such collective maintenance of the external prerequisites of action too can be avoided. After all, hermits are defined as individuals who go it alone, who refuse to interact with, let alone accept the assistance from others, and who, to use the language of this essay, actively avoid engaging in any form of second-personal address in the robust sense employed here. Such lives might turn out to be pretty solitary (poor, nasty, brutish, and short), but conceptually, it seems, there is nothing incoherent about envisioning such a scenario, and our intuitions seem to indicate that it remains compatible with agency and practical identities being present therein.

On closer analysis, however, it turns out that a hermit society very much involves instances of second-personal address, and not merely "from within" the hermits' first-personal stance. Admittedly, the conception of "address" in play here needs to be a broad one, but this can be accommodated without undermining the plausibility of the general point. When our hermits practically deliberate, when they set ends, and perform actions, then this entire process is permeated by mutually acknowledged practices of non-interference. Since, as was argued above, maintaining any such practices is something that agents must do together, even hermits have to acknowledge those cognitive second-personal prerequisites that underlie their ability to do exactly that. Even if the practices that enable hermits to go about their business uninterruptedly have nothing to do with cooperation, mutual assistance, etc., they nevertheless actually address one another in the way that the Neo-Kantian criticism seems to regard as merely optional. Taking up the second-person standpoint, then, does not require the explicit utterance and presentation of demands—Korsgaard *et al.* are right about this. However, demands on one another are made, and because of our unique practical capacities (the inescapable "freedom" to subject (or to not subject) one's behavior to a particular regime of public norms that guarantees non-interference), we cannot presume our first-person standpoints to be the result of actions that take place in a "default setting" of assured spheres of individual agency. In the case of us humans, non-interference is an artificial (though normatively non-optional) practice

and not a "natural baseline" that falls into place by itself without being in need of any shared acknowledgment of second-personal dispositions—not even amongst hermits.

To my mind, the places in *The Second-Person Standpoint* where Darwall appreciates the role of shared non-interference practices the most are the passages in which he discusses Fichte's principle of right.[13] Darwall sums up "Fichte's point" as claiming that "second-personal engagement invariably" requires all involved parties to recognize "spheres of freedom within which individuals have enforceable rights to do as they will and with respect to which others are required to forebear interference" (Darwall 2006, 262). The suggestion presented in this section is that the form of second-personal engagement just mentioned is itself "invariably" taking place as soon as one human acknowledges the sheer presence of another *qua* the unique challenge of interference that the other human is. This is true of hermits too, then, who deliberately turn their backs on one another and go it alone. Hermits, *qua* successfully constituted agents, are committed to Fichte's principle of right and to granting minimal spheres of non-interfered-with action and agency even if they do not take up the second-person standpoint in the more robust way that we, non-hermits, take for granted when we deal with others.

Ultimately, the normative force of Fichte's principle originates in the inescapable task of having to constitute oneself and others in a particular way, a way that is non-optional for beings like us, whose non-interfered-with actions are not to be taken for granted as self-evident phenomena, but are in need of shared normative presuppositions. Asking for some extra reason for taking up the second-person standpoint (that a well-unified agent can endorse/reject), therefore, is a malformed question. Since, in order to have any reasons at all even hermits must coherently structure their deliberate standpoint (as Korsgaard insists), which in turn is conditional on engaging in non-interfered with action, searching for apparent reasons (second- as well as first-personal) that an agent has to have in order to tackle the human plight gets the task of self-constitution and identity-formation the wrong way around.

13. Fichte's principle of right is: "I must in all cases recognize the free being outside me as a free being, i.e., I must limit my freedom through the concept of the possibility of his freedom" (Fichte in Darwall 2006, 262).

IV. *Why Second-Personal* normativity?

Towards the end of the previous section it was claimed that hermits, *in so far as they acknowledge the other humans around them as agents*, are drawn into the second-personal realm of making and acknowledging claims on one another's conduct and will. Even if a hermit attempted to avoid any engagement whatsoever with other agents, the dependence of her own agency on minimal practices of non-interference renders this attempt futile. Individuals who acknowledge one another as nothing more than hermitic agents nevertheless address one another in the second-personal mode: Once others are acknowledged as agents and, hence, as second-personally competent creatures who are capable of subjecting practical deliberation and choice to principles and reasons, even a hermit's normative self-conception incorporates the stance of taking up the second-person standpoint regarding those others. Each hermit does claim and demand that others at least abstain from interference with her self-constituting activities: even a hermit's practical identity non-optionally requires the issuing of such claims and demands in the presence of other agents.

Another objection, pertaining to the issue of mutually acknowledged agency, must be confronted at this point.[14] So far the power of the claim concerning the non-optionality of the second-person standpoint isn't yet fully unconditional. The issue is the assumption concerning the hermit case just summarized: The hermits acknowledge one another as normatively competent agents. The objection we turn to now calls into question the inescapability (and hence non-optionality) of having to acknowledge the others around oneself as agents in this sense. The objector asks: Even if it is granted that self-constituting action (and, hence, self-constituted agency) depends for its possibility on a minimum level of external and environmental predictability and stability, how does this fact get us into the realm of normativity, let alone closer to a successful argument for the non-optionality of the second-person standpoint? Even if this is empirically and psychologically highly implausible, why cannot we treat the other humans around us in an "unacknowledged way," the objector continues, and in the same way in which we treat other "forces of nature" such as tornados and non-human animals? The latter phenomena too

14. I am indebted to an anonymous referee for presenting this objection in an enormously clear and helpful manner. I've been confronted with that kind of objection on several occasions and try to address it more comprehensively in Hanisch 2013.

present threats of interference to our self-constitution, but, due to their partly predictable nature and behavior, they do allow some level of self-constituting action to be successfully achieved and maintained without at all introducing the notions of action, agency, and interpersonal normativity. Hence, why cannot we adopt the same stance towards other human beings, incorporate them into the causal network of psychologically and empirically predictable natural forces and events, and regard them as "metaphysical zombies" as opposed to agents who are subject to reasons? It might well be that *once* I have acknowledged other agents as individuals with normative competences, I will non-optionally find myself within the realm of second-personal demands and claims (concerning non-interfered with agency at the very least); the hermit case represents exactly this scenario.[15] But, and this remains the core of the objection, why is that prior and initial acknowledgment of other humans *qua* agents, as opposed to forces of nature, itself inescapably implicated in *my* self-constitution and agency in the first place?

There are many avenues available to respond to this line of objection. One promising response rests on the controversial, but at the same time long and widely discussed, paradigm that finds its most-prominent manifestation in Kripke's engagement with the late Wittgenstein's remarks on the possibility of a private language (Kripke 1982). More or less close precursors to this view are found in Aristotle's, Hegel's, and Anscombe's accounts of action and agency. Put very bluntly, and this is all that can be accomplished here, this paradigm is skeptical with respect to the possibility of what one might call "solitary and solipsistic normativity." A

15. On this point I therefore disagree with the second line of argument incorporated in the referee's objection. That line of argument claims that my view, since it reduces the second-person standpoint to merely regarding others as "being subject to reasons," cannot account for the defining feature of that standpoint that Darwall considers crucial, viz., that it is a perspective under which we "make and acknowledge claims on one another's conduct and will." Hence, the objector concludes, even if my account were successful in rendering the acknowledgement of others *qua* agents non-optional, this would still fall short of providing the basis for the inescapability of second-personal address in Darwall's more demanding sense.

According to the argument in the text, on the other hand, the acknowledgment of others as being subject to reasons, *in conjunction with* the fact that my own self-constitution is dependent on those others not interfering with my actions, *results* in the mutual issuing of the claim and demand to abstain from such interference. As argued in section III, in virtue of successfully acting in the presence of other acknowledged agents, I acknowledge, at that very moment, the latter not merely as creatures with the normative competence of reasons-responsiveness, but as being committed to making exactly those kinds of claims that have to do with our "conduct and will" in Darwall's sense.

born Robinson Crusoe[16] who always and necessarily populates the universe alone cannot count as subjecting himself successfully to normative standards and principles since the latter feat is achievable only within a community of "rule followers" who are defined by their entanglement in shared practices and customs. Such a community collectively upholds the "correctness standards" that determine what the adequate interpretation of the normative requirements in question consists in and, hence, whether or not a particular requirement has been satisfied in an individual agent's action and thought.

According to my interpretation and application of the Wittgenstein-Kripke paradigm (and closely related to the above-discussed reflections by Fleischacker), other independent agents and wills, who join in shared normative practices, are required for any kind of normativity to take on a conclusive and stable form—a form unachievable for lone Crusoe and his self. The experience of confronting other independent agents, who generate and maintain the normative environment for each individual ("correctness standards"), is not merely an additional and dispensable feature of agency, i.e., an optional "bonus" that renders instantiations of the latter merely less solitary, poor, nasty, brutish, and short. The dependence of our own *normative* self-conceptions on the presence of other independent and normatively active creatures goes much deeper than this. The more or less predictable forces of nature alone, on the other hand, are not enough to guarantee all of the external prerequisites for successful self-constitution then. In order to see this more clearly, we must again have a look at what exactly it is that is constituted in the course of individual action.[17]

Keep in mind that the self-constituting activities that we are discussing amount to self-constitution into an agent and, hence, into a creature that occupies *normative* standpoints—a perspective from which she deliberates,

16. The notion of a "born Crusoe" and similar devices feature in a number of texts and thought experiments concerned with the issue of solitary rule-following. Born Crusoe appears in Blackburn 1984 (who acknowledges Michael Dummett for introducing this character).

17. We must ignore for the purposes of our discussion the possibility that certain higher order animals share, to a certain degree, our normative capacities and, hence, could help born Crusoe to achieve and to maintain his agency. More complex are the objections that come from an individualistic and Kantian direction and that consider an individual's internal cognitive resources, such as the capacity to comply with the two imperatives of practical reason, not only necessary but also sufficient for successful self-constituting action. Of course, all this leads us back to the Hegelian criticisms of Kant and the more recent debate between communitarian and liberal conceptions of the self. I cannot solve these disputes here, nor can I do this with regard to the Wittgenstein-Kripke paradigm. The point of introducing these arguments is to clarify what is at stake in the currently discussed objection and what issues a possible rejoinder has to focus on.

chooses, and acts in accordance with normative principles and standards (i.e., rule-like normative entities) that she regards herself as being subject to. If it is the case that the very normativity (like every normativity, according to the currently presumed paradigm) of the principles that constitute such a standpoint is dependent on the presence of other agents (not zombies, because they cannot be members of a rule-*follower* community), then the option of treating others as mere forces of nature remains closed off. Human beings *qua* individual agents (as opposed to mere biological organisms) are indeed dependent on the presence of other agents then and cannot but acknowledge these others as normative creatures and, hence, at least as minimally second-personally competent.

Put differently, agents cannot, *pace* the currently discussed objection, avoid actually acknowledging other "species members" as agents (in a practical, not biological, sense of "species") once the opportunity to do so presents itself. This is so because their own agency's need for normativity requires such an acknowledgment of and by others. A born Crusoe, who will never encounter other beings that are capable of subjecting themselves (and others) to reasons, does in fact not constitute himself successfully into an agent, even if he might well continue vegetating as a member of the biologically-defined species *homo sapiens*; biological existence and leading a life as an agent are not the same thing. Admitting that some kind of behavior is possible in the case of Crusoe does therefore not imply that his agency and the related normative self-conception (Crusoe's practical identity) are possible, let alone comprehensible to us. As Charles Taylor summarizes, in expanding on a famous remark of Aristotle's, "as humans this separation [between individual and society; C.H.] is unthinkable. On our own, as Aristotle says, we would be either beasts or Gods" (Taylor 1985, 8).

Of course, the success of this line of argument is conditional on the adequacy of my contestable interpretation and further development of the controversial Kripke-Wittgenstein paradigm concerning the impossibility of solitary and solipsistic normativity. If Blackburn (1984) is right when he claims that a lonely born Crusoe is perfectly well capable of subjecting himself to exclusively self-imposed and self-maintained normative principles in the same way in which real-world and social agents are, then acknowledging others as normative creatures too turns out not to be a necessary condition of individual agency. This would be so, again, because if born Crusoe were able to successfully constitute himself into an agent by merely looking at the world in terms of causal and psychological laws and

regularities, then the same perspective would have to count as sufficient in a world in which other humans are present (even if this is empirically and psychologically highly implausible[18]). And if that were the case, the inescapability of the second-person standpoint too would again become questionable. I, on the other hand, tried to substantiate the case that, firstly, human beings are not automatically agents; secondly, that agency always means *normative* agency (conceiving oneself as subjecting oneself to practical rules and principles, etc.); and, thirdly, that the normativity of my agency's principles is dependent on other independent agents (not merely forces of nature) being around.

Above, it was emphasized that mutual non-interference amongst humans, conceived as unique (since capable of practically deliberating) sources of interference, is an actively upheld practice, the collective maintenance of which requires all affected parties to presuppose in themselves and others the shared capacities to so regulate their behavior. We now submitted an attempt to establish the non-optionality not merely of these mutually acknowledged non-interference practices but of acknowledging others as having the capacities to acknowledge and uphold the norms and rules that constitute such practices. Rejecting the imperatives of non-interference undermines the prerequisites of the very process that is necessary for taking up a coherent first-person standpoint to begin with. This is the interpersonal side of the "human plight" of reflective self-consciousness that consists in the inescapability of having to choose what incentives to take as grounds for action in the presence of others.

Conclusion

The difficulties related to establishing the non-optionality of taking up the second-person standpoint are due to there being two related, but nevertheless distinct, levels involved in the above line of argument, and it has not always been easy to clearly identify them as such. There is, on the one hand, the idea that human agents, in so far as they acknowledge

18. This qualification concerning the implausibility of regarding others as "metaphysical zombies" is of course merely an empirical conjecture and fails to address the objection's core thesis in a philosophically powerful way. Still, it is worth reemphasizing that even the objector grants that once others *are* acknowledged as agents (something "anthropologically necessary") this gets us into a good position to argue for at least a qualified non-optionality of Darwall's picture, regardless of whether or not the communal view of rule-following is tenable.

one another as such, conceive of human beings as a distinct and unique source of interference. They regard themselves and those around them as being able to deliberately abstain (and fail to abstain) from interference with one another and, hence, as capable of regulating the threat of the mutual annihilation and cancellation of their agency by means of subjecting their actions to shared practices and norms. The first element of my defense of the second-person standpoint has been to insist that the task of action and self-constitution in the presence of others so conceived confronts *all* agents with a unique challenge, related to the inescapable nature of the "human plight." Second-personal address amongst such potential interferers is implicated in every instance of successful self-constituting action (even amongst hermits), simply when they mutually acknowledge their presence and grant one another assured free spheres of agency.[19] It is the coherence and stability of one's very own first-personal deliberative standpoint (which especially the Neo-Kantian critics of Darwall identify as non-optional) that ultimately accounts for the non-optionality of together upholding shared norms of non-interference. The possibility of the latter in turn rests on the acknowledgment of others as possessing second-personal capacities and competencies.

The second level of the main argument had to deal with the supposed optionality of acknowledging others in exactly the sense assumed on the first level. Developing a thought put forward by Kripke's Wittgenstein, it was argued that it is the individual's own first-person standpoint that requires the acknowledgment of all others in this particular way of presupposing their second-personal competence. The need to acknowledge others as agents is due to the nature of the object of our self-constituting activities, i.e., our *normative* and *practical* self-conceptions. We must acknowledge others as agents (as opposed to more or less predictable forces of nature)

19. In a recent paper, Darwall considers the question of what it means "for two or more people to be with one another or together." And his answer, though it is awarding a much more prominent role to emotions such as empathy, somewhat resembles the answer suggested above: "Two people are with one another or together in the relevant sense, when, in one another's company or presence, they relate to each other or sense their mutual willingness to do so along with their mutual awareness of this mutual willingness. People who are thus together or with one another are open to one another and mutually aware of their mutual openness" (Darwall 2011, 6). In yet another recent paper, Darwall makes some similarly relevant claims concerning the way in which any exercise of second-personal competence lends warrant "for a demand *not to usurp others'* moral *agency* or otherwise undermine the conditions of moral choice" (Darwall 2010a, 43; my emphases). I believe, though these issues cannot be taken up here, that these developments of the second-person standpoint account might turn out to be congenial to the social extension of the self-constitution paradigm defended in this paper.

because the normativity of those principles and reasons that constitute our practical identities is conditional on other agents upholding, together with us, shared standards and structures of adequacy and correctness regarding thought and action—a solitary born Crusoe necessarily fails at this task. Furthermore, the Wittgenstein-Kripke paradigm, on which this point about acknowledging others as agents rests, suggests that Darwall is right to emphasize that this presupposition commits all agents to acknowledging a certain, minimal, conception of equality, that regards everyone as passing a minimal threshold of normative competence (otherwise we would not be able to maintain the normativity-guaranteeing practices and customs together). However, and this has been one of the main replies to the Kantian view that "a second person within" is sufficient for practical identity (in the presence of others), it is at least equally important to constitute the other agents that surround oneself as independent from one's own perspective and volitional constitution. And it is this independence that cannot be "simulated" in purely solipsistic deliberation and internal "dialogical" engagement with one-self.

Acknowledgements

I would like to thank the participants at the workshop on Stephen Darwall's work at the University of Vienna for very valuable comments and suggestions. Research for this paper was funded by the ERC Advanced Grant "Distortions of Normativity." I want to thank the project's leader Herlinde Pauer-Studer, and Alexandra Couto and Veli Mitova for extremely helpful comments on an earlier draft of this paper.

References

Blackburn, Simon 1984: "The Individual Strikes Back". *Synthese* 58, 281–301.

Darwall, Stephen 2006: *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

— 2007: "Reply to Korsgaard, Wallace, and Watson". *Ethics* 118, 52–69.

— 2010: "Reply to Schapiro, Smith/Strabbing, and Yaffe". *Philosophy and Phenomenological Research* LXXXI, 253–264.

Darwall, Stephen 2010a: "Moral Obligation: Form and Substance". *Proceedings of the Aristotelian Society* CX, 31–46.
— 2011: "Being With". *The Southern Journal of Philosophy* 49, 4–24.
Enoch, David 2006: "Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action". *Philosophical Review* 115, 169–198.
Fleischacker, Sam 2009: "Review of Stephen Darwall's *The Second-Person Standpoint*". *Utilitas* 21, 117–123.
Hanisch, Christoph 2013: *Why the Law Matters to You*. Boston, MA: De Gruyter.
— 2015 [forthcoming]: "The Hegelian Solution to the Paradox of Self-Constitution". *Hegel-Jahrbuch*.
Korsgaard, Christine M. 1996: *The Sources of Normativity.* New York, NY: Cambridge University Press.
— 2007: "Autonomy and the Second Person Within: A Commentary on Stephen Darwall's *The Second-Person Standpoint*." *Ethics* 118, 8–23.
— 2008: *The Constitution of Agency. Essays on Practical Reason and Moral Psychology.* New York, NY: Oxford University Press.
— 2009: *Self-Constitution: Agency, Identity, and Integrity*. New York, NY: Oxford University Press.
— 2011: "Natural Goodness, Rightness, and the Intersubjectivity of Reason: Reply to Arroyo, Cummiskey, Moland, and Bird-Pollan". *Metaphilosophy* 42, 381–394.
Kripke, Saul A. 1982: *Wittgenstein on Rules and Private Language*. Oxford: Basil Blackwell.
Pauer-Studer, Herlinde 2010: "The Moral Standpoint: First or Second Personal?". *European Journal of Philosophy* 18, 296–310.
Schapiro, Tamar 2010: "Desires as Demands: How the Second-Person Standpoint Might Be Internal to Reflective Agency". *Philosophy and Phenomenological Research* LXXXI, 229–236.
Taylor, Charles 1985: *Philosophy and the Human Sciences. Philosophical Papers 2*. New York, NY: Cambridge University Press.

Grazer Philosophische Studien
90 (2014), 131–147.

KANT AND THE SECOND-PERSON STANDPOINT

Jens TIMMERMANN
University of St Andrews

Summary
This paper examines the Kantian credentials of Stephen Darwall's theory of moral obligation. Discussions of duties to oneself, beneficence and gratitude, the relation between claims and duties, respect and self-conceit, the rational authority of moral requirements and the justificatory strategies in the *Groundwork* and *Critique of Practical Reason* strongly suggest that Kant would reject an account of obligation in second-personal terms. His moral theory is first-personal and should not be considered a precursor of contractualism. Emphasis on obligation notwithstanding, Kant's ethics can be read as an example of an older tradition that is centred on the agent's worth and character.

This paper examines the Kantian credentials of Stephen Darwall's second-personal theory of moral obligation. Darwall seeks to align himself with the Kantian cause, using notions like respect for the dignity of persons, the formality of the moral law, the fact of reason and autonomy as building blocks for his own contractualist theory. I shall argue, however, that he relies on an—admittedly widespread—streamlined and modernised version of Kantian ethics. Kant would have rejected the idea that obligation is essentially interpersonal; his theory of obligation is fundamentally and radically first-personal. That is why Kant's ethics should not be considered a precursor of contemporary contractualism. Emphasis on obligation notwithstanding, it can be read as an intriguing example of an older ethical tradition that is centred on the agent's worth and character.

1. *Whatever happened to duties to oneself?*

One reason why Kantian morality is not second-personal stems from those duties that human beings have towards themselves. Readers of the

Groundwork[1] are familiar with two standard examples: the narrow duty not to throw away one's life and a wider duty to cultivate one's natural talents, as derived from various versions of the categorical imperative. But duties to oneself are explained in much greater detail in the Doctrine of Virtue of the *Metaphysics of Morals*, where they gain even more prominence: moderation in matters of food and drink, self-knowledge and even truthfulness are all discussed under this heading.

As the name indicates, duties to oneself are irreducibly first-personal. They need not involve any person other than the agent. Moreover, they solely depend on the agent's will and cannot be externally enforced. I may, of course, have to perfect a talent to fulfil some social role, but that does not render the general duty to develop one's natural gifts other-regarding or interpersonal. As such, duties to oneself are not, in Kurt Baier's words, the moral community's 'business' (Darwall 2006, 27). Members of the community may in their minds approve or disapprove, but it is not their job to praise or censure. Other things being equal, I am not accountable to anyone other than myself for whether I live up to my duty to myself. Nor do these duties arise as I recognise the legitimate claims of others. They are not second-personal.

Note that the exclusion of duties to oneself from the moral realm—be it explicit or by implication—cannot be justified as a mere classificatory exercise. It is impossible to pick and choose Kantian duty types à la carte. Both duties to oneself and duties to others follow from the same moral law. Both rely on the notion of universality[2], and both can be expressed in terms of the status of human beings, oneself and others, as ends in themselves. On several occasions, Darwall quotes the full 'Formula of Humanity', which enjoins us to use humanity*, whether in our own person or in that of any other,* always as an end and never merely as a means (e.g. Darwall 2006, 229, my italics; cf. G, IV 429). But he does not comment on the idea of humanity in one's own person, an idea that may well cause complications for using Kant's ethics to explain moral obligation in second-personal terms.

---

1. Quotations from Kant's works are adapted from Kant 1996 and Kant 2011. As is customary, citations refer to the (roman) volume and (arabic) page numbers of the Berlin Academy edition, reprinted in the margins of most other editions and translations. I use the following abbreviations for individual works: G = *Groundwork of the Metaphysics of Morals*; C2 = *Critique of Practical Reason*; MM = *Metaphysics of Morals*; MM2 = *Moral Mrongovius II*.

2. If in a slightly different way in that duties to oneself depend on temporal universality only (maxims must be able to be adopted consistently at all times) whereas duties to others rely in addition on the thought of universal acceptance by all moral agents. On the idea of temporal universality as underlying duties to oneself see Glasgow 2003.

What is more, Kant repeatedly tells us that, of the two types of duty, duties to oneself take pride of place. They are more fundamental. In the Doctrine of Virtue, right at the outset of the main body of the text (§§ 1–3), he defends the idea that without duties to oneself 'there would be no duties whatsoever', not even 'external duties', i.e. duties of performance, specifically to others (MM, VI 417). This is an extraordinary thesis, and there is no denying that the argument used to establish it is enigmatic.[3] But at the very least, the opening pages of the last of Kant's three great works on moral philosophy demonstrate that he accorded duties to oneself a pivotal place in his theory of moral obligation.

So, if Kant is right—if duties to oneself are a prominent, even foundational type of moral obligation, and if they do not involve the second-person standpoint—moral obligation cannot be essentially second-personal. Or, conversely: any theory of moral obligation that is essentially second-personal (that cannot ever be merely first-personal) will struggle to account for duties to oneself and thus to claim Kant as an ally.

2. *Imperfect duty, second-personal claims and gratitude*

There is another argument from the scope of moral obligation that puts pressure on Darwall's second-personal theory of moral obligation. It concerns the nature of (as Kant variously calls them) 'imperfect', 'wide', 'non-essential' or 'meritorious' duties to others. The worry is that a theory of moral obligation based on second-personal *claims* might not be able to account for such duties, or for the characteristic response on the side of the recipient, namely gratitude. My argument takes Kant as its starting point, but the point I wish to make in this section is more general. It does not depend on narrowly Kantian premises.

Kant provisionally adopts the standard rationalist distinction between perfect and imperfect duty, but he never makes good on his promise to provide a systematic classification of duties (cf. G, IV 421 fn.). As a result, the exact nature of imperfect duties—particularly the 'latitude' that they permit—remains controversial in Kantian ethics to the present day.[4] Yet some features of imperfect duties are clear. Unlike a perfect duty, a rule of imperfect duty does not generate obligations whenever it applies. It

3. I try to reconstruct it in detail elsewhere. See Timmermann 2013a.

4. For views at opposite ends of the demandingness spectrum see, e.g., Baron 1995 and Hill 2002.

generates 'grounds of obligation' or 'obligating reasons' (cf. MM VI 224), not all-things-considered token obligations.[5]

Being helpful is an imperfect duty, but there is no actual obligation to help everyone who needs assistance because assisting must, in various ways, be possible.[6] To be obligatory, an act of assistance must be called for, i.e. there must be someone who needs my help. It also must be morally possible: providing assistance must not run counter to the commands of perfect duty, or indeed to any other competing moral claims I judge to take precedence. Moreover, I should not go so far in assisting others that I eventually need their assistance (MM, VI 454). Complete selflessness can be contrary to duty. In this sense imperfect duties are not unconditional.[7] Token obligations can be difficult to predict. That is why Kant sometimes succumbs to the temptation of reserving the honorific title of moral 'law' to perfect duties or duties of right, demoting the prescriptions of imperfect duties or duties of virtue to the level of mere 'rules'.[8]

Why do imperfect duties sit uneasily with second-personal claims? Because for Kant rightful second-personal claims are indigenous to the sphere of strict or juridical duty, at least if moral obligation is meant to have those normative implications that, following Mill, Darwall intends them to have: second-personal accountability, in particular censure, punishment or blame in the case of non-compliance as the fitting reaction to a 'wrong' (e.g. Darwall 2006, 26f., 92f., 224). For Kant, failure to comply with an identifiable imperfect obligation to others[9] does not, ceteris paribus, warrant any of these reactions. I do not wrong the stranger I refuse to help. Nor does the stranger have the authority to demand my help. Unlike perfect duties, imperfect duties are not 'owed' (G, IV 429).[10] Crucially, it does

5. I discuss the nature of obligating reasons and the question of how they can come to conflict in Timmermann 2013b.

6. Like his predecessors, Kant thought such limitations could not arise with regard to perfect duty. Perfect duties are unconditional requirements of omission with which we can always comply, see Timmermann 2013b, 45.

7. Of course, they are still *motivationally* unconditional, i.e. categorical, in that they do not rely on the agent's pre-existing ends.

8. See MM2, XXIX 619f.

9. Let us assume that this is a simple case in which considerations of friendship, gratitude, contractual obligation, promises etc. play no role.

10. This is partly a function of the indeterminacy of imperfect duty. Note that, by contrast, I do wrong a person I gratuitously hurt or harm, and my not doing so is something he can legitimately demand. Such actions make me a *bad* or *wicked* person, which is not the case if I merely refuse to help. People who fail to take their obligations of assistance seriously lack virtue. They are not as *good* as they ought to be.

not follow from any of this that I may not have a distinct, unconditional non-negotiable token obligation to help the stranger. Circumstances permitting I do. And every moderately virtuous person will help. If I do not help I am not evil or vicious, but I am not as good as I ought to be.[11] The crucial point is that when I decide to comply with an imperfect obligation I must do so, in a strong sense, of my own accord. That is why I acquire some 'merit' with regard to the beneficiary (cf. G, IV 430), and why he owes me gratitude in return.[12]

It is not difficult to see why Darwall's second-personal theory of moral obligation might seem to undermine the possibility of moral gratitude. Why should anyone be grateful to a benefactor who is merely *responding* to pre-existing second-personal claims, who (what is worse) is accountable for his actions to an external authority and who (worst of all) faces sanctions if he fails to comply?[13]

Ultimately, the reason why Kant's ethical duties to others are not second-personal is that they are generated by considerations of coherence with regard to the agent's own willing. Take the fourth and final example used to illustrate the workings of the categorical imperative in the *Groundwork* (G, IV 423). I notice someone I can easily help. At first my reaction is determined by the inevitable tendency to use what is mine exclusively for my own purposes. However, pure practical reason reminds me that it is impossible to will the selfish[14] principle I am naturally tempted to act on as a universal law. Such a law would thwart my natural egoism, which commits me to accepting assistance when I need it. At this point the equal status or dignity of all autonomous human beings does enter the moral equation, if not in its second-personal form. Rather, it makes me see that if I were justified in adopting a maxim of selfishness I would have to grant, on pain of contradiction, that everyone else facing the same choice would also be right to adopt the same maxim, which is precisely what I cannot coherently want. In this way, the duty of beneficence is the result of first-personal reflection.

---

11. Assuming my failure does not result from a principled decision *never* to help; that *would* be morally vicious (MM, VI 390).

12. Again, no one has an obligation to be grateful for mere non-maleficence.

13. Darwall briefly discusses gratitude: 'We are appropriately grateful when people benefit us or act as we wish when we lack any relevant claim or expectation of them.' (Darwall 2006, 73) But, surely, we ought to be grateful to someone who helps us in need? Which I take to be a matter of obligation, not just action on 'moral ideals' (Darwall 2006, 95).

14. Of course, my natural purposes need not be 'selfish' in the narrow sense. But they are all based on inclination, on what I would like to do.

In conclusion, Kant does not share the Millian conception of morality in which it is linked with social sanctions and blame, first because duties to oneself are essentially private and, secondly, because wide duties to others—which are not optional or supererogatory, even if actual obligations are somewhat unpredictable—involves an element of first-personal willing that cannot be cashed out in terms of responding to other people's claims.

3. *Kant on the primacy of duty (vis-à-vis rights)*

Darwall's theory of second-personal claims is reminiscent of the sphere of strict duties we owe to one another, in particular: of duties of right. (Note that Fichte makes his point expressly in terms of law or 'right'.) But even within these narrow limits Kant might disagree: like duties of virtue, duties of right rest on first-personal, not on second-personal foundations.

Contemporary political philosophy and activism tend to assume, for the most part implicitly, that rights are metaphysically and epistemically prior to duties. If you have a right to my ϕ-ing I have a duty to ϕ, and my duty to ϕ is seen as a consequence of your right. This model parallels Darwall's theory of moral obligation as based on second-personal claims. But there is reason to believe that Kant rejects the assumption of the priority of the rights of others. Duty can be understood as primary throughout his moral philosophy if individuals have rights in so far as, and because, they can claim the specific stringent obligations others have to them.

An argument for the primacy of duty can be found in a short episodic paragraph in the introduction to the Doctrine of Right.[15] Why, Kant asks, is the doctrine of morals 'usually (particularly by *Cicero*) entitled a doctrine of *duties* and not also of *rights*'? After all, there are rights that correspond to duties. The reason (he says) is this:

> We know our own freedom (from which all moral laws, and hence all rights as well as duties proceed) only through the *moral imperative*, which is a proposi-

15. Ludwig 2013. I go beyond Ludwig's account in emphasising that it is not just the corresponding duty that constitutes the right-holder's right. The right-holder must also be in a position to *claim* the right. This is manifest in the case of 'formal' duties to humanity as such, e.g. the duty not to torture. A torturer has forfeited his right not to be tortured—he is in no position to claim that being tortured in turn would be wrong. Yet torturing anyone, even torturers, is a wrong done to humanity as such. It does not depend on the torturer's personal right (now defunct) not to be tortured. That is why it is wrong to torture torturers.

> tion commanding duty, from which subsequently the capacity for obligating others, i.e. the concept of [a] right, can be devised. (MM, VI 239)

Rights can be used to make others do what we want them to do. In that sense rights obligate others. But they are made possible by the categorical imperative, which can be enforced only in so far as it concerns the equal external liberty of everyone.[16] Consider Kant's distinction between ethical and juridical legislation:

> Ethical legislation (even if the duties might be external) is that which *cannot* be external; juridical legislation is that which can also be external. (MM, VI 220)

Our rights with regard to others consist precisely in their having a corresponding duty of right that we are in a position to claim; and we know what rights we have only once we know about other people's duties. Ethical laws—despite the fact that they do, of course, require the performance of certain actions—essentially concern the agent's character and motivation. That is why they cannot be externally enforced. By contrast, limiting cases like equity and necessity aside (cf. MM, VI 233–236), compliance with juridical laws can and ought to be externally enforced. To use Kant's own example, it is an external (juridical) duty to keep a promise made in a contractual situation, and this duty can be enforced by juridical means; but 'the command to do this merely because it is a duty'—the requirement to act *from* duty and not merely in accordance with it—'belongs to internal legislation alone' (MM, VI 220). Moreover, the 'recipients' of juridical duties can be clearly specified. In that juridical duty differs from the external aspects of ethical duty to others, e.g. beneficence. A person's right is thus dependent on the fact that others have distinct and enforceable duties towards him. Strict right—comprising rights that can be properly claimed because they can be legally enforced—'is founded […] on everyone's consciousness of obligation according to the law' (MM, VI 232). The concept of duty is therefore both epistemically and metaphysically primary.

16. That is why Kant came to reject the notion of 'active' obligation (one person obligating another) as a *façon de parler* around the time he published the *Groundwork*; see MM2, XXIX 612. Obligation by another—not, ultimately, by one's own self—would count as heteronomy.

4. *Respect, self-love and self-conceit*

For Darwall, recognition respect for persons is essentially second-personal. It is 'an attitude toward individuals, not just toward a fact about a quality in them' (Darwall 2006, 131). Moreover, he finds the 'seeds' of his view in what Kant says about respect in the *Groundwork*, the second *Critique* and the *Metaphysics of Morals*. The problem with his thesis is that Kant's basic theory of recognition respect is not about others at all. It is not about individuals, nor even about other persons in the abstract.

To see why this is the case consider the phenomenology of respect for the law, which Darwall examines in some detail (Darwall 2006, 134–140). Respect for the moral law arises when the immoderate claims of sensibility are swept aside by that law in the judgement of pure practical reason. We are awed when we realise that reason has the power to do this to our sensuous nature. This reverence for the law, whose authority we come to appreciate, then serves as the motive ('incentive', *Triebfeder*) of morally good action, i.e. of action that is done simply for the sake of obeying the moral law. That is why respect is explicitly equated with 'moral interest' in doing an action (G, IV 401 fn.).

Complications arise because the claims of sensibility can be manifested in two distinct ways: as 'self-love' (*Selbstliebe, Eigenliebe*)[17] and as 'self-conceit' (*Eigendünkel*). The former is just restricted by the moral law whereas the latter is struck down altogether.[18] How do self-love and self-conceit differ? According to Darwall, self-love is the attitude of a naïve deliberator who takes subjective considerations to be objectively significant. Like a naïve experiencer, 'who takes an apparently bent stick in water to be really bent, a naïve agent may take his desire's object to be a source of reasons' (Darwall 2006, 134).

This is a good reconstruction of Kant's theory of self-love as long as we acknowledge two points. First, self-love inevitably dominates the first

17. It is not clear whether Kant intends there to be any systematic distinction between the two words. Sometimes *Selbstliebe* seems to be pre-reflective whereas *Eigenliebe* appears to occur at the level of deliberation. It is regrettable that Gregor tends to use 'self-love' for both words (Kant 1996, 199). There is, however, reason to doubt whether on this occasion Kant was aiming for terminological precision. Taking his cue from Baumgarten he advances many distinct notions of *Eigenliebe* (*philautia*)—as well as *Eigendünkel* (*arrogantia*)—throughout the years in his lectures on ethics.

18. Darwall says that 'respect' restricts self-love and strikes down self-conceit (Darwall 2006, 134), but this is not quite correct. *The moral law* restricts self-love and strikes down self-conceit (C2, V 73). *Respect* for that law arises as a result of the latter.

stage of human deliberation; i.e. it is not the case that some human beings are 'naïve agents' while others are more sophisticated. Secondly, 'taking as a source of reasons' does not amount to a final, objective judgement. Self-love does not claim full normativity. As we saw in our discussion of the Kantian duty of beneficence in section 2 above, inclinations invariably have the first word. They prompt us to be active. Kant makes this point again and again throughout his ethical writings.[19] Subsequently, as conflicting inclinations urge us to pursue their respective ends, practical reason begins to evaluate its options in terms of prudential concern, i.e. our own long-term well-being. It is hard to see how at this stage reason could claim objectivity, i.e. claim all-things-considered goodness or rationality that others would have to acknowledge. As yet, there is no acknowledgment of others as persons. Moreover, there is no practical objectivity without the moral law, which has not yet entered the deliberative process. This happens only at stage three, when judgement detaches itself from prudential deliberation to acknowledge the force of the pure practical reason.[20] If prudential concern and morality conflict, the agent faces a choice of either acting on self-interest (his own happiness) or for the sake of the (impersonal) moral law.

The natural tendency of *self-love* manifests itself at the second level. It needs to be limited to a reasonable degree by the moral law. The character flaw of *self-conceit* can appear for the first time only at stage three. It is a kind of 'arrogance', a perversion of objective standards, of being pleased with oneself and one's choices when there is no justification for it (C2, V 73). Note, however, that even self-conceit is first-personal. Self-conceit does not *rely* on second-personal considerations[21] even if, undeniably, it affects the agent's attitude towards others because a fallacious claim to objectivity implicitly amounts to a (false) expectation that others appreciate his decisions.[22]

---

19. See, e.g., C2, V 146.

20. These stages are made explicit in the 'calcareous earth' example, see Section 5 below.

21. The (erroneous) view that Kantian self-conceit turns on one's relation to others is widely held amongst Kant's interpreters. Mary Gregor's translation is once again to blame. In her words, pure practical reason strikes down self-conceit, as 'all claims to esteem for oneself that precede accord with the moral law are null and quite unwarranted' (cf. C2, V 73; Kant 1996, 199). This makes it sound as if the agent were claiming esteem *for* himself *from* others. But Kant simply talks about 'claims of self-esteem' (*Ansprüche der Selbstschätzung*), an expression that is unambiguously first-personal.

22. See Moran 2014 for a detailed discussion of these issues.

Kant's phenomenology of respect is therefore purely first-personal. It is founded on the contrast between a lower, prudentially minded self and the higher self of pure practical reason, not on the opposition between one's self and the selves of others. Self-love is limited and self-conceit struck down when I am tempted to throw away my life in exactly the same way as when I am about to make a false promise to obtain a loan.

Why, then, does Kant 'mark off instances of appraisal respect' (Darwall 2006, 132) for others in his example of the upright commoner who has greater moral merit than the decadent nobleman (echoing Fontenelle, see C2 ,V 76f.)? Because we need to distinguish between respect for the law and respect for persons. They serve different functions. The primary notion is respect for the moral law as discussed above. It is the paradigmatic moral motive. This is respect for the law within me, which I recognise as authoritative by virtue of my own autonomy. By contrast, respect for persons is a derivative notion that relates to my own agency at best in a roundabout way. Respect for persons is a form of appraisal respect:

> All respect for a person is actually only respect for the law (of righteousness etc.) of which he gives us the example. (G, IV 401 fn.)

Respect for others gains prominence in Kant's moral theory only within his philosophy of moral education by means of—mark the last word!—example. We catch a glimpse of this in the *Groundwork* and the Doctrine of Method of the second *Critique*. Even if, like the man in the gallows case (C2, V 30), we are not sure what we are going to do, we know what we ought to do, and we feel admiration for those who have done the difficult deed. That something *has been done* as a matter of historical fact—that people like Hans and Sophie Scholl defied tyranny while most of their fellow students did not—serves to show that it *can be done*. It instils in us respect for the moral fibre of those who live up to the moral law. We find it inspiring. This strengthens our own resolve to act likewise.

Of course, there are 'duties of respect' to others further down the line in Kant's moral system. They exhibit structural analogies to other kinds of respect. Also, we feel compelled to ascribe to others equal moral status as fellow 'ends in themselves', as autonomous beings equipped with transcendental freedom, which marks out persons among all other entities in the world (called 'things', *Sachen*). Without this assumption historical examples would not serve to demonstrate that moral action is practicable. (They would not help us realise our freedom.) But Kant does not usually

make this point in terms of respect, and it does not rely on relationships with other individuals.

5. *Kant on the 'overridingness' of moral requirements*

Another important point in which Darwall's understanding of morality differs from Kant's own conception concerns the notion of 'overridingness'. As Darwall notes, this is the idea that 'moral obligations always give agents conclusive reasons for acting that outweigh or take priority over any potentially competing considerations' (Darwall 2006, 26). But Kant's version of the authority of morality is much stronger.

Overridingness in the above sense is compatible with the idea that non-moral considerations have normative force regardless of their moral status. This is actually suggested by the metaphor of their being 'outweighed'. There are weights on both scales; one of them is heavier. According to Kant, however, it is not just the case that moral considerations always take normative precedence over conflicting non-moral considerations. They completely undermine any claims of competing considerations to rationality or value.

To continue the metaphor, in cases of conflict reason takes all non-moral weights off the scales. Facing an obligation to ϕ it is not just the case that I have *less* reason to follow inclination (when I cannot do both). There is, we might say, objectively no reason to follow inclination at all. This does not render inclination unattractive in felicific terms. It is not in that sense that it is 'silenced'. That is why I may well *decide* to violate my obligation to ϕ despite the fact that I know it to be supremely authoritative.[23] Objectively, in the eyes of reason, there is nothing whatsoever to be said in its favour.

This extremely strong notion of the authority of morality is at work in the opening comments of the *Groundwork* on the unconditional value of a good will. But it is spelt out most clearly in the 'calcareous earth' example of the *Critique of Practical Reason:*

> When an analytical chemist adds alkali to a solution of calcareous earth in hydrochloric acid, the acid at once releases the lime and unites with the alkali, and the lime is precipitated. In just the same way, if a man who is otherwise honest (or who just this once puts himself only in thought in the position of an honest man) is confronted with the moral law in which he cognises the

23. We shall return to this point in Section 7 below.

> worthlessness of a liar, his practical reason (in its judgement of what he ought to do) at once abandons the advantage, unites with what maintains in him respect for his own person (truthfulness), and the advantage, after it has been separated and washed from every particle of reason (which is altogether on the side of duty), is weighed by everyone, so that it can enter into combination with reason in other cases, only not where it could be opposed to the moral law, which reason never abandons but unites with most intimately. (C2, V 92f.)

Pure practical reason normatively eliminates competing inclination-based considerations. On other occasions, these considerations may well count as perfectly good reasons for action.[24] No action that contradicts morality can be in the least rational and, as reason determines goodness, no action that contradicts morality can to any degree be good. That is why Kantians are not happy with the language of reasons, which dominates the contemporary debate. The terminology of reasons suggests that certain practical considerations—say, my well-being or happiness—by themselves carry independent normative weight. But on the Kantian picture no consideration other than morality has rational weight as such. Whether a consideration can count as a reason is not a foregone conclusion. The question is whether it harmonises with the requirements of the moral law.

6. *Groundwork III: Kant's vindication of morality*

Let me add two quick notes regarding Darwall's interpretation of the vindication of morality in *Groundwork* III (Darwall 2006, 222–229). First of all, Kant's strategy differs significantly from Darwall's reconstruction. According to Darwall, Kant's deduction of morality depends on our ineliminable commitment to practical freedom, discussed in Subsection 2 of *Groundwork* III. In conjunction with the 'reciprocity thesis', which in Subsection 1 established that freedom and the moral law are two sides of the same coin, our commitment to the freedom of rational deliberation would be tantamount to an implicit commitment to the moral law in all our practical thought. Darwall correctly observes that it is implausible to assume that practical deliberation as such should commit us to the extremely strong notion of freedom that is autonomy.

24. Note that the only two factors that can determine the will are self-interested assent based on inclination and respect for the moral law.

But whatever the shortcomings of *Groundwork* III, Kant never intended the thesis that we cannot but act under the idea of freedom to bear the weight of the deduction. As commentators frequently observe, that would lead to the curious result that the main business of that section is over at G IV 448, rendering 15 large Academy pages superfluous. The most plausible reconstruction takes the argument of Subsection 2 of *Groundwork* III to make the point that we ascribe freedom to ourselves as beings that possess pure practical reason, thus prefiguring the half of the 'circle' (G, IV 450), which concludes that we are free on the basis of our consciousness of moral commands. Real progress is made only once we learn that we escape the vicious circle when we realise that we are members of an 'intelligible world'; this happens much later, in Subsection 3. Crucially, very much in the spirit of Darwall's criticism, the spontaneity of judgement is insufficient to attribute autonomy to ourselves. Rather, it is the absolute spontaneity that is manifested in our radical independence from sensibility, the capacity to generate pure ideas (G, IV 452). The actual 'deduction' takes place in the second paragraph of Subsection 4, at G, IV 454. In Subsections 5 and 6 Kant acknowledges the limitations of the project.

Secondly, what exactly *was* Kant's project in the final section of the *Groundwork*? Once again, the answer is hardly obvious. Darwall makes it look like the very strong justificatory undertaking of arguing for the validity of morality on morally neutral grounds. But it is not unlikely that even the Kant of *Groundwork* III would dismiss such a project as both overambitious and unnecessary. It is overambitious because it cannot be done. It is unnecessary because Kant assumed, for better or worse, that human beings are inevitably committed to moral standards anyway—there is no need to argue them into morality on the basis of shared non-moral assumptions because moral assumptions are already shared (if not universally enacted). The way forward with those who are found morally lacking is the moral education of respect for examples as outlined in Section 4 above. This is precisely the kind of treatment to which Kant exposes the famous 'scoundrel' at G, IV 454f. In true Enlightenment fashion, Kant thinks that deep down inside even the worst want to be good.

What, then, about the worry that morality is a 'phantasm' (*Hirngespinst*), the problem that *Groundwork* III is meant to dispel? It reflects the doubts of a decent person committed to the principles of morality but uncertain of their possibility because the source of their authority, unlike the threats of self-interest and physical determinism, is far from obvious.

Again, pure practical reason is revealed to be grounded in an intelligible world. It is on this assumption that Kant finally explains, in the deduction in Subsection 4, *how* categorical imperatives are possible as synthetic principles a priori. Kant's aims in *Groundwork* III are thus ethically more modest than Darwall makes it seem, but his strategy is metaphysically more ambitious.[25]

7. *Critique of Practical Reason: the Fact of Reason*

For whatever reason, Kant probably came to believe that the deduction of the categorical imperative in *Groundwork* III was unsuccessful. In the *Critique of Practical Reason* he expressly denies that a deduction of the principle of morality is either possible or necessary (C2, V 47). Under the heading of the Fact of Reason he now places even greater emphasis on the moral convictions of ordinary, non-philosophical moral agents. In fact, ordinary moral consciousness now serves to deduce freedom in turn.

I shall not discuss the difficult doctrine of the Fact of Reason or Darwall's second-personal reconstruction in detail. Rather, taking up Darwall's account of the 'gallows example' (Darwall 2006, 236–239, cf. C2, V 30), I should like to highlight what seems to me to be another intriguing difference between much of contemporary Kantian moral philosophy and Kant's own ethical theory. As Darwall notes, the example divides into two parts. It is used to show that freedom is not revealed by prudential deliberation—even that there *is* no freedom in prudential deliberation, which is mechanical goal-directed activity—but only by the authority of moral consciousness, when prudence and morality come apart. When a tyrant tries to bully me, on pain of execution, into giving false testimony against 'an honest man'[26], 'whom he would like to destroy', I judge that I ought to refuse, and therefore that I can act contrary to my strongest natural desire, in short: that I am free (C2, V 30).

What is curious about Darwall's interpretation of the example is that he makes it revolve around deliberation rather than action. Kant's sense of 'can', he writes, 'is simply that of an open deliberative alternative, that is, something such that one's abilities and opportunities with respect to

25. I argue for this reading of the project of a deduction in Timmermann 2007.
26. Or for that matter Anne Boleyn, his wife, cf. C2, V 155.

it do not preclude intelligible consideration of whether to do it' (Darwall 2006, 240). But a 'deliberative alternative' is not enough. It is perfectly possible to take something to be an intelligible alternative when in fact one is unwilling or unable to do it, e.g. because the will is determined by natural causes and thus not free. What Kant intends to show is not just that we can intelligibly entertain the thought that we will defy the tyrant but that we can actually do it. For Kant—but maybe not for Darwall?[27]—these two things are emphatically not the same. For Kant, morally relevant action is not determined by deliberation. Deliberation leads to viable alternatives that need to be *chosen* by a will that is free.

Note that the thesis that reasoning leads to action, not just to a clear sense of the respective merits of available options, seems to commit us to Socratic intellectualism, which makes moral failure a case of defective cognition. Bad things are done not because of a choice of the will but because we mistakenly judge the wrong option to be right, and that is due in turn to our not having a clear sense of what the right option is. Intellectualism reduces wickedness to stupidity, rendering clear-headed wrongdoing impossible and, arguably, human responsibility along with it. But this was not Kant's view, as the second part of the gallows example demonstrates. There the agent has a *very* clear sense of what he must do. If nevertheless he saves his own skin it is due to an act of will, not to further deliberation about alternatives. He acts for the sake of his own interest, and in that sense his action is perfectly intelligible. What ultimately cannot be explained is his decision to spurn the moral law.

8. *Conclusion*

In a brief reflection on the differences between competing moral theories, Darwall dissociates Kantian ethics from a fundamentally first-personal tradition that revolves around the notion of virtue:

> I believe that the role of second-personal attitudes and the second-person stance in mediating (mutual) accountability in Kantian and contractualist ethical conceptions marks a deep difference with the ethical views (frequently

27. 'When we hold people responsible, we imply that they had it within them to act as they should, not just in the sense that the alternatives were open to them or that they weren't physically prevented, *but that there was a process of reasoning* they could have engaged in *by which* they could have held themselves responsible and *determined themselves to act as they should have*.' (Darwall 2006, 241, my italics)

> ethics of virtue) of thinkers like Plato, Aristotle, Hume, and Nietzsche […], for whom evaluation of conduct and character does not take a fundamentally second-personal form. (Darwall 2006, 77)

But we have seen that Kant is very much a member of that older, virtue ethical tradition, even if Kantian virtue is to be spelt out in terms of (imperfect) obligation. For Kant as for Aristotle, the dividing line in ethics is not that between oneself and others, but the distinction between two aspects of the human self, in Kantian terms: between inclination and practical reason. For both, ethics is about the stance we take towards our natural desires. The chief difference is that the doctrine of the mean is not intended as a practical criterion of good and bad action, whereas the categorical imperative is.[28]

Postscript

This paper is largely identical with the paper I read in Vienna in March 2013. One of the most surprising and philosophically exciting results of our discussions was that the obligating second person, in Stephen Darwall's sense, need not be an actual separate person. It can be an authority within the agent's own self. This would help us account for the notion of a duty to oneself, which was the worry raised in Section 1 above. In fact, there is now a sense in which Kant's theory is second-personal after all. In his discussion of conscience as an inner tribunal in the *Metaphysics of Morals*, Kant notes that conscience

> has this peculiar feature that, although its business is a business of a human being with himself, he nevertheless views himself as necessitated by his own reason to conduct it as at the bidding *of another person*. (MM, VI 438)

A wholly unitary, undivided self would be incapable of issuing obligations. As this insight was the product of our discussing the original paper, rather than reading the book, it seems preferable to present the paper more or

28. I should like to thank Bettina Schöne-Seifert and her colleagues at the Centre for Advanced Study in Bioethics at the University of Münster for their friendship and hospitality in 2012–13. Most of the work on this paper was done during my fellowship year at the Centre. I am also indebted to an anonymous referee for *Grazer Philosophische Studien*, who encouraged me to clarify several important points. Finally, thanks are due to Herlinde Pauer-Studer and Christoph Hanisch, who organised the Vienna workshop, and to Stephen Darwall, for taking the time to discuss his work with us so fruitfully for two entire days.

less as it was when it sparked off the discussion. I very much hope Stephen Darwall will explore these matters in his own contribution to this special issue.

Literature

Baron, Marcia 1995: *Kantian Ethics Almost without Apology*. Ithaca: Cornell University Press.

Darwall, Stephen 2006: *The Second-Person Standpoint.* Cambridge, MA: Harvard University Press.

Glasgow, Joshua 2003: "Expanding the Limits of Universalization: Kant's Duties and Kantian Moral Deliberation". *Canadian Journal of Philosophy* 33, 23–48.

Hill, Thomas E., Jr. 2002: "Meeting Needs and Doing Favors". In: Thomas E. Hill Jr. (ed.), *Human Welfare and Moral Worth: Kantian Perspectives*. Oxford: Clarendon Press, 201–243.

Kant, Immanuel 1996: *Practical Philosophy*. Cambridge, MA: Cambridge University Press. Translation by Mary Gregor.

— 2011: *Groundwork of the Metaphysics of Morals: A German–English Edition.* Mary Gregor & Jens Timmermann (eds.). Cambridge: Cambridge University Press.

Ludwig, Bernd 2013: "Die Einteilungen der *Metaphysik der Sitten* im Allgemeinen und die der *Tugendlehre* im Besonderen". In: Andreas Trampota, Oliver Sensen & Jens Timmermann (eds.), *Kant's 'Tugendlehre'*. Berlin and Boston: de Gruyter, 59–84.

Moran, Kate A. 2014: "Delusions of Virtue: Kant on Self-Conceit". *Kantian Review* 19, 419–447.

Timmermann, Jens 2007: *Kant's 'Groundwork of the Metaphysics of Morals': A Commentary*. Cambridge: Cambridge University Press.

— 2013a: "Duties to Oneself as Such". In: Andreas Trampota, Oliver Sensen & Jens Timmermann (eds.), *Kant's 'Tugendlehre'*. Berlin and Boston: de Gruyter, 207–219.

— 2013b: "Kantian Dilemmas? Moral Conflict in Kant's Ethical Theory". *Archiv für Geschichte der Philosophie* 95, 36–64.

Grazer Philosophische Studien
90 (2014), 149–168.

CONTRACTUALISM AND THE SECOND-PERSON MORAL STANDPOINT*

Herlinde PAUER-STUDER
University of Vienna

Summary
This article explores Darwall's second-personal account of morality, which draws on Fichte's practical philosophy, particularly Fichte's notions of a *summons* and principle of right. Darwall maintains that Fichte offers a philosophically more appealing account of relations of right than Kant. Likewise, he thinks that his second-personal interpretation of morality gives rise to contractualism. I reject Darwall's criticism of Kant's conception of right. Moreover, I try to show that Darwall's second-personal conception of morality relies on a Kantian form of contractualism. Instead of accepting Darwall's claim that contractualism depends upon a second-personal account of morality, I will argue that contractualism provides the foundations not only for second-personal moral relations, but also for first-personal moral authority.

1. *Introduction*

The basic idea of contractualism is that moral principles are justified by a reasonable agreement between equals entertaining cooperative relations with one another.[1] Accordingly, actions are right or wrong depending upon whether they comply with principles which everyone could reasonably accept or, rather, which cannot reasonably be rejected. Contractual-

* Thanks to Fabienne Peter for helpful written comments on an earlier version of the paper. I also thank Sorin Baiasu for critical discussion. Research for this paper was funded by the ERC Advanced Research Grant "Distortions of Normativity".

1. Note that I adopt the familiar distinction between contractarianism and contractualism. Contractarianism considers an agreement on moral principles as the outcome of compromises by rational egoists who are eager to avoid suboptimal outcomes generated by their individual maximizing strategies. Contractualism assumes that agreement on moral principles is based on considerations that are acceptable from the perspective of all; no one can reasonably reject those principles.

ism is commonly associated with a relational conception of morality. A key element is recognizing the rightful claims others have on us and our accountability to them for our actions and practices.

In his book *The Second-Person Standpoint* (2006), Stephen Darwall endorses such a form of contractualism. His argument is that a second-personal theory of morality gives rise to a version of contractualism that involves Kant's requirements of universality and of treating others as ends. A striking feature of Darwall's account of morality is its reliance on Fichte's practical philosophy. According to Darwall, Fichte's conception of right, which is based on Fichte's notion of a summons, offers a better starting point for a second-personal, and thus contractualist theory of morality than Kant's practical philosophy (Darwall 2014).

This paper defends Kant's framework. Kant, as I will argue, presents a more compelling justification of a rightful condition than Fichte. Moreover, Kant's account of the normative foundations of the principle of right is, as I try to show, best understood in terms of contractualism. An implicit appeal to contractualism seems also present in Kant's ethical theory. Kant's idea of a moral community as "a realm of ends", that is a "systematic union of various rational beings through common objective laws" (Kant 1996b, 4:433, 83), can be interpreted as giving rise to contractualism. My thesis is that such a Kantian form of contractualism provides a better foundation for a second-personal account of morality than Fichte's notion of a summons and conception of right. Against Darwall's claim that contractualism relies on a second-personal account of morality, I argue that it is contractualism that provides the foundations for a second-personal standpoint in morality. Finally, I try to show that the proposed version of contractualism allows us to spell out the relations between second-personal and first-personal moral authority in the proper way. The account offered thus meets Darwall's requirement that the second-person standpoint includes first-personal considerations.

To avoid misunderstanding: The interpretation I propose amounts to a revisionary argument, suggesting that Kant's conception of morality, particularly his understanding of the constitutive principles of a moral community, aligns with contemporary versions of contractualism. While a full elaboration and defense of Kantian contractualism is beyond the scope of this paper, I try to show that an agreement-based reading of Kant's moral philosophy offers the resources for current attempts to reconstruct morality as a relational enterprise, involving reciprocal claims and obligations.

The paper is structured in the following way: After outlining (section 2) why Darwall thinks that Fichte's but not Kant's account of right supports a second-personal interpretation of morality, I argue (section 3) that Darwall is mistaken in rejecting Kant's conception of right. Section 4 points to problems in Fichte's justification of a rightful condition, and section 5 tries to show that a contractualist reading of the basic principles of Kant's practical philosophy provides the normative basis for Darwall's second-personal account of morality.

2. *Darwall's second-person standpoint and Fichte's concept of a summons*

At the core of Darwall's account of morality are four interrelated notions: claim, accountability, second-personal reason, and second-personal authority. The second-person moral standpoint presupposes that free and rational agents have second-personal authority, second-personal competence, and an obligation of accountability to others (Darwall 2006, 74ff.).[2] The validity of claims addressed to another person depends upon whether one has the legitimate authority to hold the other accountable. Second-personal relations give rise to second-personal reasons that are agent-relative. A form of reciprocal respect is part and parcel of all second-personal reason-giving. The accountability requirement is met by the "no-reasonable-rejection" test (Darwall 2006, 301).

Darwall thinks that the notion of *summons* (*Aufforderung*) as it occurs in Fichte's philosophy of right provides a model for explicating second-personal moral interaction. A summons is a second-personal claim that presupposes "a mutual second-personality that addresser and addressee share and that is appropriately recognized reciprocally" (Darwall 2006, 21). A summons, Darwall argues, leads to the recognition of oneself and the other as agents with equal normative standing.[3] He then follows Fichte's suggestion that this requires that agents are to be connected by relations of right.

The reason Darwall draws on Fichte's philosophy of right and not on Fichte's ethics, the *Sittenlehre*, in order to explicate his second-personal

2. For Darwall, second-personal address is connected with reactive attitudes like resentment, blame, indignation, and guilt. He considers these reactive attitudes as indicators of what can be rightfully demanded of others. They are the correct response if others do not recognize the legitimacy of certain claims.

3. For Darwall, the perspective of "unsummoned agency" is the observer's perspective.

conception of morality is that Darwall interprets the second-person moral standpoint as providing a foundation for contractualism. Principles of right, he argues, are crucial for contractualism: "It is a hallmark of contractualist theories that they hold principles of right to have a distinctive *role*, namely, as mediating relations of mutual respect" (Darwall 2006, 301). And, he adds, contractualism "maintains that the *form* of principles of right is mutual accountability to one another as equal persons" (Darwall 2006, 301).

Darwall's paper in this volume (Darwall 2014) further indicates his reliance on Fichte. He claims that, compared with Kant's explication of right, Fichte offers "a potentially superior" account since, unlike Kant, Fichte emphasizes the second-personal character of rights and the second-personal authority on which they are based. More specifically, while Fichte associates a right with a summons and thus with a direct way of addressing another person, Kant defines a right as the authorization to use coercion. Thus a right for Kant allows one person to treat another in a way which is according to Darwall entirely different than being involved in a second-personal normative relation *to* the other person. Moreover, he thinks that the relational obligation *to* the holder of the right to non-interference is missing in Kant's account. The person, addressed by the right holder, must respond directly to the claim of the right holder; she must recognize that she has a duty to the right holder (Darwall 2014, 12).

3. *Kant on rights and coercion*

How should we assess Darwall's thesis that Fichte offers a more plausible explication of right than Kant?

The similarity between Fichte's Principle of Right and Kant's Universal Principle of Right is obvious. Fichte's principle reads: "*I must in all cases recognize the free being outside me as a free being, i.e. I must limit my freedom through the concept of the possibility of his freedom*" (Fichte 2000, 49). Kant's Universal Principle of Right states: "Any action is *right* if it can coexist with everyone's freedom in accordance with a universal law, or if on its maxim the freedom of choice of each can coexist with everyone's freedom in accordance with a universal law" (Kant 1996a, 6:230, 387). Both principles are standards for regulating our relations in the sphere of external freedom, relying on the same idea: equal freedom is constitutive for rightful relations. Equally close are some of Fichte's and Kant's explications

of the concept of right. While Fichte holds that "the concept of right is the concept of the necessary relation of free beings to one another" (Fichte 2000, 9), Kant describes right as "the sum of the conditions under which the choice of one can be united with the choice of another in accordance with a universal law of freedom" (Kant 1996a, 6:230, 387).

These similarities notwithstanding, Darwall dismisses Kant's notion of a right. As indicated, his objection is that Kant's definition of a right in terms of authorized coercion legitimizes a certain way of dealing with the other person but does not involve a second-personal relation and "a relational obligation *to* the right-holder that is entailed by and correlative to the claim right he holds" (Darwall 2014, 12).

I think that Darwall's critique rests on a misunderstanding. It is true that Kant associates the concept of right with "an authorization to use coercion" (Kant 1996a, 6:231, 388). Darwall assumes that this authority plays out directly in the interaction of agents and thus amounts to the right of one agent to coerce another. However, Kant's definition of a right, as presented in the introduction to the *Doctrine of Right*, is not meant in that sense. Later chapters in the *Doctrine of Right* make clear that Kant attributes the authority to use coercion to the state. The right to hinder a hindrance to freedom is the state's prerogative.

As his discussion of property shows, Kant distinguishes carefully between a provisional possession of an object and *a right* to the possession of an object. An initial or original acquisition of an object is simply a claim on an external thing as one's own, thus amounting to a "provisional possession" of external objects. Kant notes that we have to leave the state of nature (where we have provisional possession of objects) and consent to "a rightful condition" of public justice that guarantees and protects property rights. Only in a state of "externally lawless freedom" would an individual be "authorized to use coercion against someone who already, by his nature, threatens him with coercion" (Kant 1996a, 6:307, 452). Kant claims that such a condition of "externally lawless freedom" has to be overcome since it is "a condition that is not rightful, that is, in which no one is assured of what is his against violence" (Kant 1996a, 6:307, 452).[4]

According to Kant, the transition to a rightful condition requires a state based upon a constitution all citizens could accept, since it secures

4. As Kant writes: "A rightful condition is that relation of human beings among one another that contains the condition under which alone everyone is able to *enjoy* his rights, and the formal condition under which this is possible in accordance with the idea of a will giving laws for everyone is called public justice" (Kant 1996a, 6:306, 450).

their rights: "Public *right is therefore a system of laws for a people, that is, a multitude of human beings, or for a multitude of peoples, which, because they affect one another, need a rightful condition under a will uniting them, a constitution (constitutio)*, so that they may enjoy what is laid down as right" (Kant 1996a, 6:311, 455). Human beings thus need a system of "public coercive laws", since in a rightful condition individuals do not have authority to use coercion themselves. Rather, they require the proper public institutions for executing coercion.

One has to be careful here: Although Kant claims that from the perspective of the normative principle of equality each individual "member of a commonwealth" has "coercive rights against every other", no individual has the right to turn that right into action since it is "the head of state, by whom alone any rightful coercion can be exercised" (Kant 1996c, 8:291, 292)[5]. Kant attributes to citizens the normative status of holding coercive rights, but he does not grant them the right to execute that coercive authority by themselves. This would represent a fall back into the conditions of a state of nature. Thus individuals will consent to transferring their coercive authority to the state. The insight that they would otherwise face a condition of "external lawless freedom" provides them with a reason to do so.

In Kant's framework, the move to a civil and rightful condition is justified since each member of the society would, if rational, consent to the normative principles of "lawful freedom". Kant's claim that a rightful condition is tied to "the idea of a will giving laws for everyone" indicates that individuals share the normative ground for obtaining a condition of public justice in which "everyone is able to *enjoy* his rights" (Kant 1996a, 6:306, 450). Kant thus seeks to outline the public normative conditions that allow for rightful interpersonal relations. Nothing rules out that those relations cover second-personal ground. Kant himself, as we have seen, points out that human beings need a rightful condition since their actions have an effect on others. Rightful relations require respecting the rights of others and include, hence, the duty of accountability. Darwall's critique of Kant's definition of a right is therefore not justified.

---

5. Compare also the following passage: "But *in terms of right* (which, as the expression of the general will, can be only one and which concerns the form of what is laid down as right not the matter or the object in which I have a right), they are nevertheless all equal to one another as subjects; for, no one of them can coerce any other except through public law (and its executor, the head of state), through which every other also resists him in like measure" (Kant 1996c, 8:292, 292).

As a matter of fact, Kant's assumptions are decisive for Darwall's own project. A second-personal account of morality implicitly presupposes a normative framework such as the one Kant has in mind when talking about a 'rightful condition'. The reason is that Darwall takes Fichte's notion of a summons as a model for explicating a second-personal way of addressing another. However, in order for a summons to be constitutive of a second-person moral standpoint, it cannot be an arbitrary kind of demand or command—a point on which Darwall agrees. Recall that he emphasizes that agents must have the *de jure* authority to make claims on another person's conduct. Without such implicit normative assumptions, a 'summons' might represent a morally inacceptable mode of relating to the other individual.

In *The Second-Person Standpoint*, Darwall discusses the case of a slaveholder addressing his slave (Darwall 2006, 267).[6] He concedes that a slaveholder's demand on his slave might just be abusive talk. While the slaveholder has authority over the slave, we would certainly deny that he has *legitimate* authority to address the slave in a way that reduces him to a mere recipient of orders.[7] Given the power relations the slave faces, he has reason to comply with the orders of the slaveholder. But this is not the kind of normative second-personal reason Darwall has in mind, indicating that not any mere summons to another person provides a basis for a second-personal account of morality. Only a summons that amounts to a second-personal address to an equal gives rise to second-personal moral relations.

Still, the question remains as to whether Fichte's emphasis on the notion of a summons does not capture more profoundly than Kant's theory what is at stake in second-personal ways of addressing each other as equals. Let us thus take a closer look at Fichte's argument.

6. Darwall discusses this case since he is aware that his position might be normatively too strong. He uses the case of the slaveholder to show that his position does not entail that bad actions or involvement in bad practices such as slavery would commit us to some sort of "pragmatic contradiction" (Darwall 2006, 265).

7. There is a certain ambiguity in Darwall's way of explaining legitimate or de jure authority. He does not distinguish clearly between legitimate or de jure authority coming with professional roles and legitimate or de jure authority in the moral sense. This is apparent when he talks about the de jure authority of a sergeant vis-à-vis her troops. The example is dazzling. Hierarchical professional roles do not generate the kind of second-personal reasons Darwall has in mind. Although we would not deny that the sergeant has legitimate or de jure professional authority to address his subordinates through orders, we might have moral reasons for rejecting her specific orders. This indicates that second-personal authority alone is not sufficient to produce second-personal moral reasons.

4. *Fichte's conception of right*

In *Foundations of Natural Right* (2000), Fichte attempts to derive the concept of right by demonstrating its indispensability to free and self-conscious agency. His idea is that an individual "cannot posit itself as a rational being with self-consciousness without positing itself as an *individual*, as one among several rational beings that it assumes to exist outside itself, just as it takes itself to exist" (Fichte 2000, 9). According to Fichte, self-consciousness involves not only the subject's awareness of herself as unifying representational states, but also the subject's practical perspective on herself as a rational and free being. Thus free and rational agency requires an external domain of freedom that is regulated by the Principle of Right. Hence, for Fichte, rights amount to necessary conditions of self-consciousness.

Fichte's deduction of the conception of right proceeds in three steps based upon three theorems.[8] The first is that a subject with self-consciousness ascribes to itself free efficacy—i.e., the capacity to form ends and express its will in the world of objects. The second step is that a subject can only see itself as having efficacy if it sees others in the same way. That is to say, a subject becomes aware of its agency via the agency of others, or more specifically, via the summons of another agent which is a call upon the subject "to resolve to exercise its efficacy" (Fichte 2000, 31). "[O]ne is driven", Fichte claims in the first corollary to this second theorem, "from the thought of an individual human being to the assumption of a second one, in order to be able to explain the first" (Fichte 2000, 38). The final step of the deduction of the concept of right is that assuming the existence of other rational beings involves standing in a particular relation to them, namely "a relation of right (*Rechtsverhältniß*)" (Fichte 2000, 39). This entails, Fichte maintains, that "*I must in all cases recognize the free being outside me as a free being, i.e. I must limit my freedom through the concept of the possibility of his freedom*" (Fichte 2000, 49).

Commentators have noted that Fichte's deduction of the Principle of Right seems problematic. The worry is that it involves an illegitimate shift from a theoretical notion of self-consciousness (the unification of object representations) to a practical form of self-consciousness, namely the willing of a self-determining agent (Neuhouser 2000, xvif.).

Indeed, the claim that rights are necessary conditions of being conscious of one's own self is hardly tenable. The thesis seems, if at all, only

8. For a helpful and clear exposition of Fichte's argument see Neuhouser (2000, xii–xvii).

plausible with respect to autonomous practical agency. In order to be free and rational agents, we need a political order guaranteeing our personal and political rights.

Fichte agrees that individuals, if they want to enjoy their rights, should enter into "a community among free beings". However, he offers only a hypothetical reason for doing so. Fichte believes it is "not possible to point to an absolute reason why someone should make the formula of right—limit your freedom so that the other alongside you can also be free—into a law of his own will and actions." He therefore attributes mere "hypothetical validity" to the Principle of Right (Fichte 2000, 82).

Darwall acknowledges that Fichte's conditional justification of the Principle of Right poses a problem for his argument that Fichte provides a more convincing account of rights than Kant does. He therefore criticizes Fichte's "voluntarism" by arguing that entering into a community of rightful relations with others should not be something an agent may or may not choose but rather a necessary normative precondition. Darwall points out that Fichte must presuppose that "[u]nless we assume that we each already have the normative standing to obligate ourselves through our reciprocal commitments, no reciprocal willing can yield any obligating law" (Darwall 2014, 18). Thus instead of voluntarism, Fichte should according to Darwall adopt a "presuppositional interpretation" of the connection between a summons and the concept of right. That means that Darwall considers a community of regulating external relations in accordance with rights granted by the principle of equal freedom as indispensable for making claims on others.

In my view, Darwall's concession that Fichte's theory has to presuppose the normative framework of a 'rightful condition' so that a summons represents a legitimate claim on another amounts to endorsing Kant's thesis that a rightful condition of public justice is a precondition for having rights towards others.[9] The Universal Principle of Right is for Kant constitutive of a normative order in which agents may enjoy their space of external freedom independently from arbitrary interventions by others.

9. One might object that Kant presupposes a natural right, namely the "innate right" of freedom. True, Kant claims that this "original right" to freedom belongs "to every man by virtue of his humanity." He adds, however, that this right holds only "insofar as it can coexist with the freedom of every other in accordance with a universal law" (Kant 1996a, 6:237, 393). This indicates that one's innate right to freedom presupposes the Universal Principle of Right and thus a rightful condition of public justice.

We can interpret Kant's point as the claim that rightful demands on others must come with a normative justification backed by principles of public morality on which free and rational agents would agree since this grants them the normative status of being respected as free agents by others. An essential principle of public morality is the Principle of Universal Right, which is the basis for "a system of laws" guaranteeing equal freedom for all. This reasoning provides the link between Kant's account of a rightful condition and contractualism. Kant uses the idea of a rational agreement for making the presupposition of a rightful condition normatively compelling.

In what follows, I aim to show that we can interpret Kant's practical philosophy as involving a form of contractualism, which provides a justification for principles of freedom on the one hand and ethical principles on the other. My argument is that Darwall's second-personal account of morality relies upon accepting such a Kantian version of contractualism.

5. *Contractualism as a basis for a second-personal account of morality*

In the final chapter of *The Second-Person Standpoint*, Darwall argues that his second-personal account of morality provides a foundation for contractualism. Here Darwall relies on Kant, not on Fichte. He offers a reformulation of Kant's central moral principles in terms of contractualism.

While I largely agree with Darwall's interpretation of Kant, I think his order of priority should be reversed: Instead of claiming that contractualism depends upon a second-personal account of morality, I argue that contractualism provides the foundations not only for our second-personal moral relations but also for first-personal moral authority.

The starting point for Darwall's reading of Kant's theory is the dignity of persons as expressed by the Formula of Humanity. This requires treating one another as ends and never merely as means. According to Darwall, the concept of dignity has to be spelled out in second-personal terms, namely those of mutual accountability among equals. Dignity thus commits us to addressing others with second-personal demands that cannot be reasonably rejected and to which free and rational agents hold themselves accountable.

Darwall maintains that the condition of recognizing others' dignity gives rise to the idea of a realm or kingdom of ends—that is, a community of rational beings united by common laws requiring us to

treat one another as ends and never merely as means. Kant's Formula of Universal Law (FUL) specifies for Darwall what this idea of a kingdom of ends entails with regard to the particular will and reasoning of the individual person. That is to say, the equal recognition of others excludes regarding individuals as having special standing— an idea that's fleshed out by asking whether one's maxims could be thought or willed as universal laws.

In short, Darwall's interpretation of Kant's framework can be expressed thus: take the Formula of Humanity (FH) as fundamental; interpret FH in terms of the Formula of the Realm of Ends (FRE); and finally, interpret the Formula of Universal Law (FUL) in light of the idea of the realm of ends (Darwall 2006, 304–309, esp. 308). The "no-reasonable-rejection" test amounts to a particular way of expressing the universalization procedure of the FUL. In other words, to ask whether my maxim can be thought or willed as a universal law is equivalent to asking whether others cannot reasonably reject actions based on that maxim.

The problem of attributing such a form of contractualism to Kant is that it seems to blur the distinction between individual and public morality. To ask which principles no one could reasonably reject, or to whose universal acceptance everyone could rationally agree, leaves open whether we are referring to ethical principles or principles of justice. Equally, the question which claims of others we cannot reasonably reject does not specify whether we should assess those demands on ethical grounds or grounds of justice. This seems to conflict with the clear line Kant draws between the spheres of internal freedom (ethics) and external freedom (justice, law).

I will now suggest a contractualist interpretation of Kant's guiding principles of practical philosophy that acknowledges the distinction between the sphere of ethics and the sphere of right. The idea is that the conception of a realm of ends, namely a community of rational agents who recognize one another as free and equal, is fundamental for Kant's ethics and his philosophy of right. Such a community involves that all its members agree on its constitutive normative principles. I then try to show that Kant's framework not only endorses a first-personal moral standpoint but can also make room for a second-personal account of morality.

Kant's clearest appeal to contractualism appears in his political philosophy. In his essay *On the Common Saying*, Kant argues that a rightful or civil condition that establishes a commonwealth preserving "the *right* of human beings under *public coercive laws"* rests on a social contract, namely "the general (united) will of the people" that "is called the *original contract*"

(Kant 1996c, 8:289, 290; 8:295, 295). The possible consent of citizens constitutes for Kant "the touchstone of any public law's conformity with right" (Kant 1996c, 8:297, 297).

Kant's ethical theory, however, seems far from contractualism. The point of Kant's argument in the *Groundwork* is to reveal the principle of a good will by a conceptual analysis of the notion of duty. This analysis leads, as we know, to the Categorical Imperative. A morally good person makes the Categorical Imperative her principle of action by acting only on maxims that can be thought or willed as universal law.

The incentive[10] of the action is decisive for the morality or immorality of the action. We act morally when we act from the motive of duty. Maxims as subjective principles of action are tied to the setting of ends. Directing one's incentives and setting one's ends is for Kant a matter of internal freedom; no person or institution has the right to force anyone else to adopt specific ends. Kant's ethics thus seems restricted to inner self-regulation and self-legislation by the Categorical Imperative. This, one might object, commits us to a first-person moral standpoint incompatible with a contractualist account of morality. The upshot of this line of criticism is that the idea of a mutual agreement on principles does not capture Kant's focus on internal incentives and maxims by assessing their moral quality.

The situation is different in the sphere of external freedom. Here what is crucial for Kant is *that* people follow the Principle of Right that obligates them to respect the equal external freedom of others. Kant considers the motivational reasons *why* persons do so to be irrelevant. Mere compliance is morally sufficient.[11] Since the sphere of external freedom does not rely on the inner incentives and motivations of the person, it seems compatible with contractualism.

How should we cope with that dividing line between Kant's ethics and his philosophy of right? Does it entail the two spheres of Kant's practical philosophy to exist side-by-side and track different theories of morality?

---

10. In the *Groundwork* Kant defines an incentive (*Triebfeder*) as a subjective ground of motivation, based on desires and inclinations, while a motive (*Bewegungsgrund*) is an objective ground that motivates a rational will. In his later works, the term 'incentive' has a broader meaning, covering empirical incentives and incentives of pure reason (Wood 1999, 111ff., 360f., note 1). This paper uses the term 'incentive' in the broader sense.

11. Kant famously expressed this distinction between internal and external freedom thus: "All lawgiving can therefore be distinguished with respect to the incentive […]. That lawgiving which makes an action a duty and also makes this duty the incentive is *ethical*. But that lawgiving which does not include the incentive of duty in the law and so admits an incentive other than the idea of duty itself is *juridical*" (Kant 1996a, 6:219, 383).

There is similarity between Kant's formulation of the Categorical Imperative in the *Groundwork* (namely to act only according to maxims which can be willed as universal law) and the Universal Principle of Right, which requires that actions be compatible with the freedom of others "in accordance with a universal law". But the exact connection remains unclear.

Actually, there seems no way to proceed directly from the Categorical Imperative in the *Groundwork* to the Universal Principle of Right. The Universal Principle of Right cannot be derived from the Categorical Imperative since the latter is tied to the motives and ends of the person, whereas the Universal Principle of Right completely ignores those internal elements. Some philosophers have thus concluded that Kant's philosophy of right does not fit into the structure of Kant's moral philosophy.[12] Kant's own project notwithstanding, the *Groundwork* does not appear to provide the foundation for Kant's moral philosophy as a whole.

My suggestion is that Kant's idea of a realm of ends, as he introduces it in the *Groundwork*, provides the unifying principle for his practical philosophy. It should be seen as the centerpiece of his practical philosophy, covering the basic principles of Kant's ethics and his philosophy of right.[13] A consequence of this view is that contractualism is the foundation for Kant's ethics and his philosophy of right.

Kant formulates the idea of a realm of ends thus: "[A]ll rational beings stand under the *law* that each of them is to treat himself and all others *never merely as means* but always *at the same time as ends in themselves*. But from this there arises a systematic union of rational beings through common objective laws […] [W]hat these laws have as their purpose is just the relation of these beings to one another as ends and means" (Kant

12. One proponent of the so-called independence thesis is Willaschek (1997) and (2009). Guyer (2009) defends the unity of Kant's practical philosophy. Ripstein (2009, Appendix) tries to explain the connection between Kant's philosophy of right and the rest of Kant's philosophy by appealing to Kant's arguments about concepts and objects in the *Critique of Pure Reason* in order to clarify why and in what respect the Universal Principle of Right has to be different from the Categorical Imperative.

13. This interpretation differs somewhat from Kant's own exposition in the *Groundwork*. But I do not think it incompatible with the spirit of Kant's ideas. At first glance, the suggestion that the realm of ends is central to Kant's practical philosophy seems to conflict with Kant's claim that the Formula of the Realm of Ends is the result of the synthesis of the Formula of Humanity and the Formula of Universal Law. However, to claim that we should relate to one another in a way that respects our being free and rational agents, as the idea of a realm of ends requires, captures the meaning of the Formula of Humanity; the idea of the Universal Law Formula is fleshed out, in my interpretation, by asking which common principles and laws can be universalized—i.e., cannot be reasonably rejected by all free and rational agents.

1996b, 4:433, 83). He then adds: "A rational being belongs as *a member* to the kingdom of ends when he gives universal laws in it but is also himself subject to these laws" (Kant 1996b, 4:433, 83).

What justifies the idea of a realm of ends, that "systematic union of rational beings through common objective laws"? One might claim that those common laws constituting a "systematic union of rational beings" are dictated by pure practical reason. However, one might also interpret them as being based upon an agreement.

We are brought directly to the idea of a realm or kingdom of ends by seeking to answer the question: On what fundamental principles must our relations to each other be based so that all of us, as free and equal agents, have reason to consent to them?[14] We would all give ourselves those common laws and choose to live by them since this guarantees our equal standing and freedom. It seems reasonable, from the standpoint of all, to accept them; we cannot reasonably reject them. This way we are a moral community, entertaining relations of dignity to each other.

The obvious next step is to argue that this general idea of a realm of ends is spelled out in the sphere of internal freedom by the ethical Categorical Imperative and in the sphere of external freedom by the Universal Principle of Right. The Categorical Imperative secures my autonomy in the sphere of inner motivations and convictions; the Universal Principle of Right warrants my independence from the choice of others, thus enabling me to be my own master in external relations to others.

Kant's practical philosophy aims to answer two crucial questions, i.e. with regard to ethics: 'What is the principle of good action?' and, as concerns the sphere of right: 'What justifies coercion?' In answering those questions, Kant offers us two regressive arguments. In the *Groundwork*, the regressive argument leads to the Universal Law formulation of the Categorical Imperative. Kant reasons that a free or autonomous will acts according to its own principle or norm, that is to say, it is guided by a self-given law. The principle of a free will is henceforth a law, and the condition of being a law, namely holding universally, is exactly fulfilled by the Categorical Imperative in the Universal Law formulation.

---

14. Even Christine Korsgaard, who defends a first-personal conception of morality, speaks the language of contractualism when she explains Kant's conception of a realm of ends in *The Sources of Normativity* in the following way: "The moral law, in the Kantian system, is the law of what Kant calls the Kingdom of Ends, the republic of all rational beings. The moral law tells us to act only on maxims that all rational beings could agree to act on together in a workable cooperative system" (Korsgaard 1996, 99).

The regressive argument in the philosophy of right is based upon the assumption that coercion is justified when it prevents an action that would violate the condition of universal freedom. As Kant puts it: "[I]f a certain use of freedom is itself a hindrance to freedom in accordance with universal laws (i.e., wrong), coercion that is opposed to this (as a *hindering of a hindrance to freedom*) is consistent with freedom in accordance with universal laws, that is, it is right" (Kant 1996a, 6:231, 388).

Kant's point is that enforceable constraints on behavior should be set by universal external laws consistent with everyone's freedom. This then grants the authority to use coercion. Crucially, this authorization amounts to a general regulation acceptable from all individual standpoints. Kant emphasizes that the use of coercion is not vindicated because of the unlawfulness of a particular act. The right to use coercion is for Kant neither directed at the inner determination of a perpetrator to comply with the external law, nor is it based upon the "unlawful use of freedom" by a perpetrator's particular criminal act. Rather coercion is warranted by universal external laws—and this universality includes the coexistence of one's freedom with the freedom of perpetrators, as Kant's remarks make clear: "Thus when it is said that a creditor has a right to require his debtor to pay his debt, this does not mean that he can remind the debtor that his reason itself puts him under an obligation to perform this; it means, instead, that coercion which constrains everyone to pay his debts can coexist with the freedom of everyone, including that of debtors, in accordance with a universal external law" (Kant 1996a, 6:232, 389).

The idea of a realm of ends and the Universal Law formulation of the Categorical Imperative and the Universal Principle of Right are connected in the following way: first, there is the contractual agreement of all subjects to the idea of a realm of ends, which includes the commitment to see oneself as belonging to a community of free and equal cooperative subjects. The regressive arguments show why the principle of ethics, the Categorical Imperative in the Universal Law formulation, and the guiding principle of the philosophy of right (i.e., the Universal Principle of Right) can be considered as implementing the idea of a realm of ends in the spheres of both internal and external freedom. I treat others as ends and not merely as means if I ask myself whether my maxims for acting can be thought or willed as universal law. As indicated, this means to ask whether others can reasonably consent to my maxim. I also treat others as ends, and not merely as means, if I consent

to live in cooperative relations with others regulated by the principle of equal freedom.[15]

On this interpretation, the regressive arguments do not simply lead to the ethical Categorical Imperative and the Principle of Universal Law—leaving the connection between ethics and the philosophy of right still open. Indeed, the regressive arguments provide a detailed account for why the ethical Categorical Imperative and the Universal Principle of Right meet the requirements set by the general standard of a community of rational beings based on "common objective laws."[16]

Before proceeding to outline the consequences of this reading of Kant with respect to Darwall's second-personal account of morality, I want to address a possible objection: Is the step from the idea of the realm of ends to the Universal Law formulation of the Categorical Imperative really plausible? In other words, does it not simply leave us again with the problem that any general principle that seeks to unite ethics and the philosophy of right ultimately fails to capture the crucial point of ethics, namely the decisive role of the incentive of action and the inner determination of the person? Can the Categorical Imperative in ethics be considered an implementation of the idea of a realm of ends?

The problem is especially relevant given that the Formula of Universal Law is addressed to the individual herself and brings her will to the fore by requiring: "*[A]ct only in accordance with that maxim through which you can at the same time will that it become a universal law*" (Kant 1996b, 4: 421, 73). Some philosophers, including Darwall, have therefore claimed that the Universal Law formulation of the Categorical Imperative makes no appeal to the standpoint of others and what they can reasonably accept

---

15. One might object that this interpretation is in tension with Kant's claim that the Universal Principle of Right is "a postulate that is incapable of further proof" (Kant 1996a, 6:231, 388). I think, however, that reconstructing the reasons we have for consenting to the Universal Principle of Right is more in the spirit of Kant's project. True enough, Kant is often close to rationalism, even a dogmatic form of rationalism. Yet his painstaking efforts in developing a regressive argument in the *Groundwork* show that Kant is not content with relying on mere a priori truth as a justification of the Categorical Imperative.

Guyer (2009, 201–217) argues that Kant's claim that a postulate is "incapable of further proof" does not mean that a postulate needs no further justification.

16. A possible criticism is that Kant does not leave room for principles of justice as standards of public morality, functioning as guidelines for the sphere of law and the legal design of the basic institutions of society. But such principles of justice could equally be reconstructed in response to the question: Which form of society would free and rational agents who want to be recognized as free and rational agents choose? Kant's position can be interpreted to cover such principles of public morality.

or cannot reasonably reject (Darwall 2006, 307). They argue that the contradiction in the case of non-universalizable ethical maxims amounts to mere self-contradiction of the inner self.[17]

Such a narrow reading of the Universal Law Formula seems to me untenable. Closer examination reveals that the universalization test only works if one assumes that others act likewise. What the inner determination of one's will amounts to is the acceptance that one's will must be governed by a principle that could be thought or willed for others as well. I have to act in a way that my will, expressed in my maxims, be guided by principles to which others could consent. We have to read this "will" as my internal voice, but not as my solipsistic voice. The decisive element in the Formula of Universal Law is universality, and this includes making my inner resolutions with regard to the standpoint of others. Given its structure, the Categorical Imperative test requires me to consider the claims of others; its application trivially presupposes something like 'second personal competence'.

The worry about an unbridgeable gap between the idea of a realm of ends and the Formula of Universal Law thus seems unsubstantiated. Coherence in my normative commitments requires that I, who already agreed on the laws constitutive for the community of rational and equal beings, approve that my own will must also be guided by those laws. I address the general principle to myself. The incentive of my action is relevant since it is indispensable to my individual agency. Moreover, when it comes to my own moral action, the incentive must be of a particular kind: I simply cannot determine myself to act morally unless my incentive is tied to the moral principle.

Individual agency in the sphere of external relations likewise requires an incentive. However, here I can, though need not act morally. This is the case since the authority for regulating the sphere of external relations is handed over to the state. And the state may require, even force us to comply with the laws, but it may not require us to do so morally.

Let us return to Darwall's theory of morality. Darwall claims that the second-person standpoint gives rise to contractualism. In outlining the connection between his second-personal conception of morality and contractualism, he focuses exclusively on principles of right. Recall his remark that principles of right constitute a "hallmark" of contractualism.

17. Ripstein (2009, 385, 386) defends such an interpretation. For him, the Categorical Imperative test "is a kind of self-contradiction for which the agent must reproach him- or herself in conscience" (Ripstein 2009, 377).

This entails that Darwall's account of morality merely captures our moral obligations in the sphere of external freedom.

But morality also includes norms that guide my moral relations to others in light of the principles to which I consented, given that general recognition of those principles secures my status as a free and rational agent. This is where first-personal moral authority becomes relevant: I consent to live by the principles constituting a community of free and rational beings, given that this grants me the recognition and respect of others as a free and rational subject. But this initial agreement on the norms constituting such a moral community entails that I, deliberating from the first-person standpoint, also assess my actions and obligations to others in light of those principles that speak to their standpoints.[18] Contractualism thus covers not only the second-person standpoint but also shapes my first-personal moral authority.

Darwall cannot—and in fact does not—dispel a first-person standpoint. An essential element in his moral theory is responsibility and accountability to others. However, Darwall himself emphasizes that this second-personal aspect must have a first-person counterpart. What he calls *Pufendorf's point* is relevant here: If we, as members of the moral community, hold another person responsible for complying with a moral obligation, we take it that the person likewise holds herself responsible. In Darwall's words:

> To intelligibly hold someone responsible, we must assume that she can hold herself responsible in her own reasoning and thought. And to do that, she must be able to take up a second-person standpoint on herself and make and acknowledge demands of herself from that point of view (Darwall 2006, 23).

This entails that the person must rely on her *own reasoning and judgment* and not simply be driven by fear of sanctions from others. Just as Pufendorf claimed that moral obligations derive not merely from the external authority of God threatening us with sanctions (in case we violate moral obligations) but from our understanding of God's demands, so too our commitment to moral obligations emerges from our understanding of the demands, which we, as rational agents, address to ourselves. To take up a second-person standpoint on oneself means to define one's first-personal moral authority in light of the principles constituting the moral community of free and rational agents. By confirming the importance of "free self-determination" (Darwall 2006, 23), Darwall presupposes a

18. For Darwall, the first-person perspective of "unsummoned agency" amounts to a mere observer's perspective on objects and alternative actions (Darwall 2014, 14).

kind of internalism on the part of the individual subject: the agent herself acknowledges the force of obligations.

Contractualism does not rule out such first-personal considerations. Even if the normative force of the basic moral laws rests on a contractualist agreement with others, there must be corresponding first-person recognition of that source of normativity.

To conclude: I argued that contractualism offers a direct route to the normative idea of a community of equals constituted by principles that cannot be reasonably rejected. However, contractualism also allows us to specify that general idea in order to make room for the crucial distinction between a theory of justice and rights, on the one hand, and ethics, on the other. Such a form of contractualism grounds Darwall's second-personal account of ethics but also covers the first-personal standpoint.

Literature

Darwall, Stephen 2006: *The Second-Person Standpoint. Morality, Respect, and Accountability*. Cambridge, MA and London: Harvard University Press.

— 2013: *Morality, Authority, and Law. Essays in Second-Personal Ethics I*. Oxford: Oxford University Press.

— 2014: "Why Fichte's Second-Personal Foundations Can Provide a More Adequate Account of the Relation of Right than Kant's" (this volume, 5–20).

Fichte, Johann Gottlieb 2000: *Foundations of Natural Right. According to the Principles of the Wissenschaftslehre*. Cambridge: Cambridge University Press. Translation by Michael Baur. References to this work are to page numbers.

Guyer, Paul 2009: "Kant's Deductions of the Principles of Right". In: Paul Guyer, *Kant's System of Nature and Freedom*. Oxford: Oxford University Press, 198–242.

Kant, Immanuel 1996a: "The Metaphysics of Morals. Part I: Metaphysical First Principles of the Doctrine of Right". In: Kant, *Practical Philosophy*. Cambridge: Cambridge University Press, 363–506. Translation by Mary Gregor. References are to the Preussische Akademie edition and, second, to the page numbers of the Cambridge volume.

— 1996b: "Groundwork of The Metaphysics of Morals". In: Kant, *Practical Philosophy*. Cambridge: Cambridge University Press, 41–108. Translation by Mary Gregor. References are to the Preussische Akademie edition and, second, to the page numbers of the Cambridge volume.

— 1996c: "On the Common Saying: That May Be Correct in Theory, But It Is of No Use in Practice". In: Kant, *Practical Philosophy*. Cambridge: Cambridge University Press, 273–309. Translation by Mary Gregor. References are to the Preussische Akademie edition and, second, to the page numbers of the Cambridge volume.

Koorsgard, Christine M. 1996: *The Sources of Normativity*. Cambridge: Cambridge University Press

Neuhouser, Frederick 2000: "Introduction to Fichte's *Foundations of Natural Right*". In: Johann Gottlieb Fichte, *Foundations of Natural Right. According to the Principles of the Wissenschaftslehre*. Cambridge: Cambridge University Press, vii–xxviii.

Pauer-Studer, Herlinde 2010: "The Moral Standpoint: First-Personal or Second-Personal". *European Journal of Philosophy* 18, 2, 296–310.

Ripstein, Arthur 2009: *Force and Freedom. Kant's Legal and Political Philosophy*. Cambridge, MA and London: Harvard University Press.

Willaschek, Marcus 1997: "Why the Doctrine of Right Does Not Belong in the *Metaphysics of Morals*: On Some Basic Distinctions in Kant's Moral Philosophy". *Jahrbuch für Recht und Ethik* 5, 205–227.

— 2009: "Right and Coercion: Can Kant's Conception of Right Be Derived from his Moral Theory?". *International Journal of Philosophical Studies* 17, 1, 49–70.

Wood, Allen W. 1999: *Kant's Ethical Thought*. Cambridge: Cambridge University Press.

Grazer Philosophische Studien
90 (2014), 169–182.

FALSEMAKERS: SOMETHING NEGATIVE ABOUT FACTS

Michele PAOLINI PAOLETTI
University of Macerata (Italy)

Summary
The author argues for the existence of negative facts. The first section is devoted to an argument, grounded on truthmaker maximalism, that aims at demonstrating that negative facts must exist at least as false propositions' falsemakers. In the second section, the author analyzes and criticizes several attempts to get rid of negative facts: the ones based on incompatibilities, absences, totality facts and polarities, as well as the ones based on various restrictions on truthmaker maximalism or on the non-acceptance of facts as truthmakers. In particular, it is shown that an ontology that accepts negative facts is simpler than an ontology that denies their existence and that in general, many attempts to get rid of negative facts turn out to recognize the existence of such entities or of entities that are more mysterious than negative facts themselves.

One of the most interesting aspects of truthmaker theories concerns the relationship between negative truths and the supposed existence or non-existence[1] of negative facts. However, the debate on negative facts is older than the one on truthmaker theory: it started with Russell's and Wittgenstein's well-known theses, respectively, on the existence of negative facts and on the non-existence of negative properties. In what follows, I shall argue that, if we accept the truthmaker maximalist principle, we have to accept the existence of both truthmakers and falsemakers and, in turn, we have to accept the existence of negative facts (§1). Afterwards, I shall analyze and criticize some attempts to get rid of negative facts (§2).

In sum, I shall try to deal with this problem:

(A) do negative facts exist? What grounds do we have for accepting their existence?

1. I assume here a general notion of existence, according to which, roughly, everything that is part of our best ontological inventory exists.

In my opinion, we should accept negative facts' existence *at least* because they play an ineliminable role as false propositions' falsemakers.

1. *The argument*

My argument for the existence of negative facts is quite simple and it develops some intuitions of Russell's (Russell 2010, 41–47) and Brownstein's (Brownstein 1973, 43f.). According to a principle (the truthmaker maximalist principle), which is accepted by some philosophers who also accept the truthmaker theory,

(TMP) a truth-bearer (e.g. a proposition) is true iff there exists something (a truthmaker, e.g. a fact) which makes it true.

I accept this principle, while other philosophers, as we will notice, deny it.

However, a fact is an obtaining state of affairs, while a truth-bearer is something which can be acceptably said to be true (or false), namely something which can bear some definite truth-value. If this principle works with the truth-value "true", why cannot we state that there is some corresponding principle which works with the truth-value "false"? This principle (the falsemaker maximalist principle) will state that

(FMP) a truth-bearer (e.g. a proposition) is false iff there exists something (a falsemaker, e.g. a fact) which makes it false.

If we accept (FMP), it seems that we have to accept the existence of negative facts too. In fact, let me consider the proposition expressed by

(1) Mario lives in France.

This proposition is false. It is false because it is true that

(2) Mario does not live in France.

By (FMP), (1) is made false by a negative fact, which makes (2) true. In other words, if (FMP) is true and (1) is false, there is some negative fact which makes (1) false. This fact is Mario's not living in France, and it is

the same fact which makes the proposition expressed by (2) true. Thus, there exist negative facts.

It seems reasonable to accept (FMP) and I do not see any argument against it. Even if one might deny that each truth has its own truthmaker (namely, that there are at least as many truthmakers as truths), why do we have to deny that each truth has a truthmaker (even if there are truthmakers grounding more than one truth)? Furthermore, the proposition expressed by (1) is false because Mario does not live in France. It follows that it is reasonable to accept that there is something which makes (1) false. If, according to (TMP), something can be said to be true in virtue of some positive fact (i.e. the instantiation of some property or relation by something), (1) can be said to be false, *per analogiam* and given the truth of (FMP), in virtue of some negative fact. Does this analogy work? And why do we have to accept (TMP) and (FMP) with regard to every true or false proposition?

2. *How (not) to get rid of negative facts*

One classical objection against the existence of negative facts was provided by Wittgenstein in his *Tractatus* (Wittgenstein 1961, proposition 5.44). According to Wittgenstein, if we accept such facts, we have to accept an infinite number of facts grounding the truth of one single proposition. Let me consider (2). The proposition expressed by (2) is made true by the negative fact that Mario does not live in France, namely that it is not the case that Mario lives in France. Yet, if there were such a fact, there would be another negative fact, namely the fact that

(2′) it is not the case that it is not the case that it is not the case that Mario lives in France.

On the other hand, if

(3) Mario lives in Italy

expresses a true proposition, which is made true by the fact that Mario lives in Italy, accepting the existence of negative facts, the proposition expressed by

(3′) it is not the case that it is not the case that Mario lives in Italy

would be made true by another fact, which would be different from the fact which makes (3) true. In sum, if we accept that there are negative facts and if we accept (TMP), following Wittgenstein, we have to admit, for any proposition P, that ~~P and P, if P is true, are made true by two distinct facts. Yet, it seems that ~~P is made true by the same fact which makes P true. Thus, there are no negative facts. Considering (FMP), if P is false, then ~~P is false too. Yet, they are made false by the same fact (and not by two different facts). However, this objection does not work with my argument. In fact, I can accept that, if P is false, P and ~~P are made false by the same negative fact, so that there need not be infinite negative facts which respectively make P, ~~P, ~~~~P, etc. false. On the other hand, if P is false, then ~P is true, and it is true in virtue of the same (negative) fact which makes ~~~P true. According to my perspective, if P is true, then there is a positive fact which makes it true and all the propositions such as ~~P, ~~~~P, etc., that are logically equivalent with P (and have the same constituents, couples of negations excluded), are made true by the same positive fact. If P is false, there is a negative fact which makes it false, and this is the same fact which makes ~P, ~~~P, etc. true. In order to deal with Wittgenstein's objection, I do not have to deny, following Grossmann (Grossmann 1992, 79f.), that such logically equivalent propositions are made true by the same fact.

Russell, in his *Philosophy of Logical Atomism*, accepts the existence of negative facts, and I shall consider one Russellian remark on negative facts when I shall deal with Raphael Demos' incompatibilism. It seems that, in order to avoid the ontological commitment to negative facts, either we have to reduce negative facts to positive ones, or we have to modify (TMP). It is possible to identify at least six different attempts to reduce negative facts to positive ones:

(p1) (1) is made false and (2) is made true by the same positive fact which makes (3) true (*naive "positivism"*);
(p2) the positive properties of living in France in (1) and of living in Italy in (3) are incompatible, so that, if the fact expressed by (3) is true, the fact expressed by (1) is false and (2) is true (*properties' incompatibilism*);
(p3) the negative fact which seems to make (2) true and (1) false does not exist by itself: it is only a negative description of a positive fact (*Demos' incompatibilism*);

(p4) (1) is false because it does not have any truthmaker (*absence solution*);
(p5) (1) is made false by some positive fact, the fact that (3), *and* by some positive totality fact, which concerns all the properties of the subject of (1) and (3) or which concerns the world (*totality facts solutions*);
(p6) the positive state of affairs expressed by (1) has polarity 0, so that it is false, and the positive state of affairs expressed by (3) has polarity 1, so that it is true (*polarity solution*).

With regard to (TMP), we might:

(tmp1) restrict (TMP) to positive truths;
(tmp2) refuse Truthmaker Maximalism, according to which every true proposition is made true by something in the world;
(tmp3) deny that facts are propositions' truthmakers.

I shall consider (tmp1), (tmp2) and (tmp3) in the last part of this paper.

The solution (p1) (*naive "positivism"*) seems to be easily understandable. The proposition expressed by (3) and which is made true by a positive fact (the fact that Mario lives in Italy) provides a falsemaker for (1). Thus, the positive fact that (3) (the fact that Mario lives in Italy) is the most suitable falsemaker for (1). It seems quite obvious that the negative fact that (2) (the fact that Mario does not live in France) is related to the positive fact that (3) and that the former fact is the falsemaker for (1). Yet, the negative fact that (2) is not contained in the positive fact that (3), since (3) does not imply (2). We have to add one further premise to (3):

(4) it is not the case that Mario both lives in France and in Italy.

Thus, if (4) and (3) are both true, they imply the truth of (2) and the falsity of (1). Yet, (4) expresses a negative proposition, which can be made true by some (particular) negative fact, until it is proved otherwise. Furthermore, material implication is not enough and there seems not to be any necessary logical equivalence between (4) *and* (3), on the one hand, and (2), on the other hand: it might be the case that (3) is false, because Mario truly lives in Germany, for example, and that (2) is true. If we substituted (4) with

(5) Mario lives in France or (*aut*) he lives in Italy,

we might establish some kind of logical equivalence between (5) and

(5′) it is not the case that Mario lives in France *and* that he lives in Italy, *but* it is the case that Mario lives in France or (*vel*) in Italy.

The first conjunct of (5′) would express a negative proposition and would be made true by some negative fact, until it is proved otherwise. If (5) and (5′) are logically equivalent, why do we have to choose (5) instead of (5′) in order to express the proposition that we must conjoin with (3)? Some philosophers might answer that an ontology which excludes negative facts is simpler than an ontology which admits them. Thus, if (5) and (5′) are logically equivalent, we have to admit just the positive fact expressed by (5). Yet, in order to get rid of negative facts, we would have to admit positive (strongly) disjunctive facts (such as the fact that (5)). Thus, this simpler ontology would have to admit simple positive facts, such as the fact that (3), *and* positive (strongly) disjunctive facts. We would both have two different categories of facts. Yet, the former ontology would have to admit, in order to get rid of negative facts and to provide a falsemaker for (1), complex conjunctive facts, such as the fact that (5) *and* (3), until it is proved otherwise, while my ontology would only have to admit the negative fact that (2). Thus, my ontology, which comprehends two categories of facts (simple positive facts and simple negative facts), would be simpler than the former ontology, which would comprehend simple positive facts, positive (strongly) disjunctive facts and complex conjunctive facts, such as the fact that (5) *and* (3). It would be possible to reply that this latter conjunction does not express any further kind of facts. Yet, in order to accept this reply, we would have to modify or to refuse (TMP). In fact, (2) expresses a true proposition, while (1) expresses a false one. Thus, there must exist at least one fact which makes (2) true and (1) false. This fact is the fact that (5) *and* (3), and it is not the fact that (5) alone *or* the fact that (3) alone. If we modified (TMP), we would have to admit that, in this case, there is something which makes (2) true *and* which is not a fact. Yet what would the truthmaker of (2) be?

I shall try to object against (p2) (*properties' incompatibilism*) in a similar way. According to (p2), (2) is made true and (1) is made false by one positive fact, i.e. the fact that (3), *and* some kind of incompatibiliy between the properties of living in France and of living in Italy. Yet, this incompatibiliy is a fact and it seems to be the fact that (5). If we accepted this fact, we would run into the same difficulties that we already met in analyzing (p1).

Furthermore, it would be possible to ask the following question: why is (5) true? If we did not want to accept the positive (strongly) disjunctive fact that (5), we would have to admit that this fact corresponds to a conjunction between some general law expressed by

(6) it is not possible to live at the same time and in the same respect in two different places

and some apparently positive fact, i.e. the fact that

(7) France and Italy are two different places.

Yet, firstly, it seems that (6) is made true by a general negative fact, until it is proved otherwise. Secondly, to state that (7) is nothing but to state that

(7′) it is not the case that France and Italy are the same place,

and (7′) would be made true by some negative fact, until it is proved otherwise. Thus, we can notice that there are propositions, such as the one expressed by (7), which are made true by negative facts, even though they seem to be made true by positive facts. In fact, there are grammatical predicates (such as the predicate "being different") that are positive, but intrinsically are nothing but abbreviations of the negation of other positive predicates (such as, in this case, the predicate "being identical"). Thirdly, it seems that the properties of living in France and of living in Italy are incompatible because the property of living in Italy, when it is truly attributed to something in conjunction with some true general law, implies the instantiation of the negative property of non-living in France. I do not think that those who do not accept the existence of negative facts would accept the existence of negative properties.

Developing one of Russell's objections, I shall now object against (p3) (*Demos' incompatibilism*). According to Demos (Demos 1917), we do not have to accept the existence of negative facts, but we have to accept that some positive facts can be described in a negative way. For example, the fact that (3) can be described with regard to its opposition to the fact that (1). Thus, we can have some apparently negative fact, such as the one which seems to make (2) true. Yet, Demos' incompatibilism is quite problematic. In fact, as Russell remarks, there cannot be incompatibility between facts. On the contrary, it seems that there is some kind of incompatibility only

between propositions: two propositions are incompatible iff they cannot be both true. And, if we accept (TMP), the truth of the proposition expressed by (3) and the truth of the proposition expressed by

(8) the propositions expressed by (1) and (3) are incompatible,

there must be some fact which makes (8) true. For the same reasons that I have already considered with regard to (p2), this fact would be an apparently positive, yet truly negative fact, because it would be the fact that

(8′) it is not the case that the propositions expressed by (1) and (3) are compatible.

Thus, Demos' incompatibilists have to accept negative facts.

However, Demos has another objection against the existence of negative facts: they cannot be experienced, while positive facts can be experienced. This is not a good objection: if we consider experience in a narrow sense, there are also positive facts (e.g., the fact that there is a positive number between 3 and 5) that we cannot experience. It is true that, when we see a red ball, we see that it is red and we do not see that it is not green. Yet, given that there are also existing positive facts which cannot be perceived, I do not think that we have to deny that negative facts exist just because we cannot perceive them. If we consider experience in a broad sense, it seems that we can know that there are negative facts, as well as we can know that there are positive facts: I can know that the ball is red, as well as I can know that the ball is not green. Furthermore, as it has been recently remarked in (Barker & Jago 2012), the perception of holes is the perception of some negative fact.

It seems to me that solutions (p4) and (p6) are less interesting than the others. According to (p4) (*absence solutions*), the proposition expressed by (1) is false and the proposition expressed by (2) is true because the former proposition does not have any truthmaker (for example, Mumford 2007). Yet, how might we express this absence? The proposition expressed by (1) has no truthmaker. What does it make this latter proposition (the proposition that [the proposition expressed by (1) has no truthmaker][2]) true? If an absence were a negative fact, it seems obvious that (p4) would not be a good solution against the existence of negative facts. This absence

2. I use square brackets in order to distinguish propositions from statements.

might correspond to or be contained within some totality fact involving Mario. Yet, as I shall try to show, solutions such as (p5) (*totality facts*) have their troubles too. An absence seems to be a mysterious entity. Thus, in this case, we would have to introduce one mysterious kind of entities (the absences) in order to get rid of another kind of mysterious entities (the negative facts), and these two kinds of entities, furthermore, seem to be suspiciously similar. Thus, in order to admit (p4), there remains that we have to modify (TMP).

The other less interesting solution is (p6) (*polarity solution*), according to which the difference between negative and positive facts is a difference in polarity (Beall 2000, Priest 2000). For example, the fact which makes (1) false and (2) true has polarity 0, while the fact which makes (3) true has polarity 1. However, polarity solution does not deny that there are negative facts. Perhaps, it seems to admit that there are facts which are made positive or negative by their polarity. Yet, at first, we do not have to confuse facts and states of affairs: facts can be considered a subset of states of affairs, namely they can be considered obtaining states of affairs. The same state of affairs can have two different polarities, while the same fact seems to have just one polarity: if this fact is negative, it has polarity 0, while, if it is positive, it has polarity 1. However, what would polarities be? The analogy between physical polarity and the polarity of facts seems not to be sufficient to provide an exhaustive metaphysical explanation. Following Dodd, it seems that this solution "merely provides us with a notation for a theory of facts, not an account of such facts' ontological nature" (Dodd 2007, 392).

According to the solution (p5), a proposition such as the one expressed by (2) is (also) made true by some "totality fact" involving the referent of its subject. Following Armstrong (Armstrong 2004, 53–60), it is possible to state that this proposition is made true by a huge conjunction of facts involving Mario and all his positive properties (every fact of this conjunction corresponds to the instantiation of some positive property by Mario) *and* a "limiting fact", according to which these are all the properties that Mario instantiates. This latter fact can be expressed in negative terms too ("it is not the case that Mario has any other positive property, except P_1, P_2, etc.", where P_1, P_2, etc. are all the positive properties instantiated by Mario). Thus, why do we have to consider this fact positive? This is a typical objection against Armstrong's thesis (Molnar 2000, 80ff.; Dodd 2007, 388f.). Furthermore, it has been argued against Armstrong that the "limiting fact" involving Mario is not one additional fact different from

Mario's positive facts (Molnar 2000, 81), that Armstrong does not obtain a minimal truthmaker for (2) and that many positive facts involving Mario are irrelevant in order to make (2) true (Björnsson 2007). For the same reasons for which I do not agree with Armstrong's thesis, I do not agree with the second of these objections: it seems to me that the "limiting fact" adds something to all the positive facts of Mario's and that it must be distinguished from them. Yet, my general objection against Armstrong's solution is that there is no reason to accept his "minimalism" with regard to the existence of facts. The conjunction between the propositions about Mario's "limiting fact" *and* all the positive true propositions about Mario (that seem to be made true by all the positive facts about Mario) is logically equivalent to the conjunction of all the true propositions about the negative facts involving Mario. Let me assume that Mario instantiates a finite set of positive properties: for example, he instantiates three hundreds properties, that I shall call P_1, P_2, … P_{300}. He does not instantiate a finite set of positive properties: for example, he does not instantiate one million positive properties, that I will call P_{301}, P_{302}, … $P_{1000300}$. Thus we will have the following situation:

(9) ((Mario is P_1 *and* Mario is P_2 *and* … *and* Mario is P_{300}) *and* (P_1 *and* P_2 *and* … *and* P_{300} are *all* the positive properties instantiated by Mario)) *iff* (Mario is *not* P_{301} *and* Mario is *not* P_{302} *and* … *and* Mario is *not* $P_{1000300}$).

Does (9) express a true logical equivalence? It seems not to do. In fact, we should add that there are no other positive properties in the world, except P_1, …, $P_{1000300}$. Thus we will have:

(10) ((Mario is P_1 *and* Mario is P_2 *and* … *and* Mario is P_{300}) *and* (P_1 *and* P_2 *and* … *and* P_{300} are *all* the positive properties instantiated by Mario) *and* (P_1, …, $P_{1000300}$ are all the positive properties instantiated in the world)) *iff* (Mario is *not* P_{301} and Mario is *not* P_{302} *and* … *and* Mario is *not* $P_{1000300}$).

Does the fact that (10) (or the facts in the left side or the fact in the right side of this logical equivalence) provide a truthmaker for the true proposition expressed by (2)? Yes, it does. Yet, it provides it iff the proposition that [(2)] is part of the huge conjunction of all the propositions involving Mario in the right side of the logical equivalence. The minimal truthmaker

for (2) will be one of these latter conjuncts, namely the negative fact that (2). In order to get rid of negative facts, we have come back to our starting point, after having introduced further complicated facts, that we do not have to introduce in order to make (2) true.

According to Cheyne and Pigden (Cheyne & Pigden 2006), we have to consider the world (or some part of the world) as it is in order to make (2) true. This "big fact" (the world as it is) makes (2) true. Yet, I do not see the advantages of this solution over Armstrong's solution. If the "big fact" is the world as it is, there are facts that are part of the "big fact" that are irrelevant in order to provide a truthmaker for (2): the facts that Paris is in France, that London is in England, and so on. The world as it is is not the minimal truthmaker for (2), as well as the conjunctive fact that Paris is in France *and* London is in England (or the set that is constituted by these two atomic facts) is not the minimal truthmaker for

(11) London is in England.

Furthermore, how should we consider this "big fact"? According to the authors, it is a big positive fact or a collection of (positive) facts. They state that, if we dismiss the thesis according to which truthmaking is something like entailment, we do not have to admit "limiting facts" (or "totality facts"): the only big fact which will provide a truthmaker for (2) will be a fact involving the world as it is. Yet, if this big fact is a fact involving the world, how can it be identical with all the positive facts involving the world's things? We should have to transform every fact about Mario into a fact about the world as it is. This is quite strange. Furthermore, if we consider the big fact no more than a conjunction of facts involving the world, I do not see how we can avoid that there is one further "limiting fact" involving the world. If the world is identical with "*all* the obtaining states of affairs", i.e. with "*all* the facts", how can we state that this definition does not contain in itself some "limiting fact" in addition to all the world's positive facts? Looking for a truthmaker for (2), we would avoid the admission of this "limiting fact" only if we admitted that the world contains negative facts too. Furthermore, the big fact is not a minimal truthmaker for (2): in order to make (2) true, we might consider only one relevant part of the world.

In reply to my objections against (p5), the following solution might be suggested: all the positive facts put together bear some kind of relation (the totaling relation) with the first-order property of being a fact. This seems

not to commit us to negative facts, since such (further) fact is a positive one. Yet, this solution states something that is simply false. It is not true that all the positive facts put together (except the totality fact) total the first-order property of being a fact, since, if there is a positive fact that makes the proposition expressed by such a statement true, that new positive fact (the totality fact) is added to all the positive facts put together. On the other hand, if we included the totality fact within the positive facts that stand together in the relation of totaling with the property of being a fact, then such a totality fact would not be a totality fact, since there would be some new totality fact (i.e., the fact that all the above considered positive facts *plus* the old totality fact total the first-order property of being a fact). Thus, it is not the case that all the positive facts put together total the first-order property of being a fact *or* the truthmaker for such a statement (more exactly, for the proposition that is expressed by it) is not a fact or it is not a positive fact.

I shall briefly consider the solutions expressed by (tmp1), (tmp2) and (tmp3). According to (tmp1), we might restrict (TMP) to positive truths. D. H. Mellor (Mellor 2003), for example, considering true negative existential statements, states that they are true because there exist no truthmakers for their negations. Similarly, the proposition expressed by (2) is true iff (1) does not have any truthmaker. Yet, the proposition [the proposition that (1) does not have any truthmaker], if it is true, is made true by something. What does it make this latter proposition true? I do not know. We might deny that there is something which makes this proposition true, given that this latter true proposition is a negative one. Yet, in this case, we would have no reasons for which this proposition is true. (TMP) expresses one important intuition about true propositions: the intuition according to which, in order for *every* proposition to be true, there must be something in the world that grounds the truth of that proposition. If we dismissed (TMP), we would have a high price to pay because there would exist true propositions whose truth would not be grounded in reality. As I have already noticed, this does not imply that each proposition must have its particular truthmaker.

According to (tmp2), *not all* the true propositions are made true by something in the world. Yet, what does it make these latter propositions true? I think that those who accept (tmp2), or something similar to (tmp2), have the *onus probandi*. The same problem arises with regard to (tmp3): if facts are not propositions' truthmakers, what are their truthmakers? I cannot examine here all the answers given to this question. In general, philosophers

invoke tropes or single objects or the world itself as an object. Schaffer (Schaffer 2010) and Cameron (Cameron 2008) argue that the world as it is can be considered a truthmaker for negative propositions. However, it is in turn a fact involving the world that the world is as it is, that it does not contain Mario's living in France and that it cannot contain it (otherwise, it would be another world). Furthermore, according to this solution, one might state that the actuality of the world (of *this* world) necessitates Mario's not living in France. Yet, is not the actuality of the world a fact?

One could reply that these objections are question-begging, since they assume that the actuality of the world (as well as that the world is as it is) is a fact. However, let me consider a possible world that is slightly different from the actual world in which Mario lives in France: let me call such a world w_2 and let me call the actual world w_1. The world w_2 makes it true that (1) and it makes it false that (2), while w_1 makes it false that (1) and it makes it true that (2). Thus, (1) is true in w_2 and it is false in w_1, while (2) is false in w_2 and it is true in w_1. However, how can it be that (1) is false *simpliciter* and (2) is true *simpliciter*? (1) and (2) have the truth-values that they do in fact have since w_1 is the actual world and w_2 is *not* the actual world *and* since we accept that whatever is true in the actual world is true *simpliciter* and whatever is false in the actual world is false *simpliciter*. In reply, one could deny this latter principle and accept that there is nothing such as truth *simpliciter* and falsity *simpliciter or* that it is only true *simpliciter*, e.g., that [it is true in w_1 that Mario does not live in France] as well as it is only true *simpliciter* that [it is true in w_2 that Mario lives in France]. Furthermore, one could deny that it is literally the case that Mario (our Mario, the one that is part of the actual world) lives in France in w_2, provided that Mario is not part of w_2, while one of his Lewisian counterparts is part of it. Finally, one could claim that *being the actual world* (or simply *being actual*) is not a genuine property of w_1, so that it is not a genuine fact that w_1 is the actual world. I cannot deepen here all the problematic aspects of such replies. I think that there is a genuine sense according to which something is true *simpliciter* and whatever is true/false in the actual world is true/false *simpliciter*. Furthermore, it seems to me that Mario himself (and not one of his Lewisian counterparts) lives in France according to w_2, and, finally, I do not see any reason to deny that the property of being actual is a genuine property of w_1.

In sum, I think that the traditional solution involving facts is the simplest one and, if we do not find out insurmountable problems, we should carry on accepting it.

Armstrong, David Malet 1978: *Universals and Scientific Realism. Volume II: A Theory of Universals.* Cambridge: Cambridge University Press.

— 2004: *Truth and Truthmakers.* Cambridge: Cambridge University Press.

Barker, Stephen & Jago, Mark 2012: "Being Positive about Negative Facts". *Philosophy and Phenomenological Research* 85, 117–138.

Beall, J. C. 2000: "On Truthmakers for Negative Truths". *Australasian Journal of Philosophy* 78, 264–268.

Björnsson, Gunnar 2007: "If You Believe in Positive Facts, You Should Believe in Negative Facts". In: Toni Rønnow-Rasmussen, Björn Petersson, Jonas Josefsson & Dan Egonsson (eds.), *Hommage à Wlodek. Philosophical Papers dedicated to Wlodek Rabinowicz.* Lund: Lund University. http://www.fil.lu.se/hommageawlodek/index.htm.

Brownstein, Donald 1973: "Negative Exemplification". *American Philosophical Quarterly* 10, 43–50.

Cameron, Ross 2008: "How to Be a Truthmaker Maximalist". *Noûs* 42, 410–421.

Cheyne, Colin & Pidgen, Charles 2006: "Negative Truths from Positive Facts". *Australasian Journal of Philosophy* 84, 249–265.

Demos, Raphael 1917: "A Discussion of Certain Types of Negative Propositions". *Mind* 26, 188–196.

Dodd, Julian 2007: "Negative Truths and Truthmaker Principles". *Synthese* 156, 383–401.

Grossmann, Reinhardt 1992: *The Existence of the World. An Introduction to Ontology.* London & New York: Routledge.

Mellor, David Hugh 2003: "Real Metaphysics: Replies". In: Hallvard Lillhammer & Gonzalo Rodriguez-Pereyra (eds.), *Real Metaphysics. Essays in Honour of D. H. Mellor.* London & New York: Routledge, 212–238.

Molnar, George 2000: "Truthmakers for Negative Truths". *Australasian Journal of Philosophy* 78, 72–86.

Mumford, Stephen 2007: "Negative Truth and Falsehood". *Proceedings of the Aristotelian Society* 107, 45–71.

Priest, Graham 2000: "Truth and Contradiction". *The Philosophical Quarterly* 50, 305–319.

Russell, Bertrand 2010: *The Philosophy of Logical Atomism*. London & New York: Routledge.

Schaffer, Jonathan 2010: "The Least Discerning and Most Promiscuous Truthmaker". *The Philosophical Quarterly* 60, 307–324.

Wittgenstein, Ludwig 1961: *Tractatus Logico-Philosophicus.* London & New York: Routledge.

Grazer Philosophische Studien
90 (2014), 183–200.

MODAL MEINONGIANISM AND CHARACTERIZATION REPLY TO KROON

Francesco BERTO
University of Amsterdam and University of Aberdeen

Graham PRIEST
CUNY Graduate Center and University of Melbourne

Summary
In this paper we reply to arguments of Kroon ("Characterization and Existence in Modal Meinongianism". *Grazer Philosophische Studien* 86, 23–34) to the effect that Modal Meinongianism cannot do justice to Meinongian claims such as that the golden mountain is golden, and that it does not exist.

1. *Introduction*

Meinongianism is the view that some objects do not exist. After some decades in the wilderness, the view is now, rightly, resuming its place on the philosophical landscape. In fact, it has been pretty orthodox for most of the history of Western philosophy: its period in the wilderness is an historical aberration (see Priest 2008). Given the developments in the techniques of logic since its falling from favour, the view can now be articulated with a precision and determination that it did not, before, enjoy. In fact, there are various such articulations on the market. In a recent paper, Fred Kroon raises interesting objections to one variation of it (Kroon 2012: page references are to this unless otherwise indicated). The point of *this* work is to reply to such objections.

The objections center around the *Characterization Principle*. This is a principle that tells us something about the properties of objects that may not exist. As a first cut, this is that an object, characterized in a certain way, has those properties it is characterized as having. It can be framed for either definite or indefinite descriptions. Since definite descriptions can be defined in terms of indefinite ones (*the so and so* is *a unique so and so*), we

will mostly employ indefinite descriptions in this essay. Using ε as such a description operator, then, when $A(x)$ is any condition with free variable x, one can understand the naïve Characterization Principle as claiming:

CP: $A(\varepsilon x A(x))$

No one, however—Meinongian or otherwise—can endorse this: in a two line argument, it leads to triviality. Let $A(x)$ be the condition $x = x \wedge B$, with arbitrary B. Let t be $\varepsilon x A(x)$. Then by CP, $t = t \wedge B$; and so, B (see Priest 2005, *viii*).

One approach to the problem is to restrict the Principle to a limited vocabulary, composed of predicates often called *nuclear* (see Parsons 1980, Routley 1980, who calls the predicates *characterizing*). Thus, CP_N is CP restricted to those $A(x)$ which contain only such predicates. A different approach, the so-called "dual copula" approach (Rapaport 1978, Zalta 1983, 1988), is to hold that $\varepsilon x A(x)$ does not instantiate $A(x)$ at all; rather, it bears the relation of *encoding* to it, $A_E(\varepsilon x A(x))$, where $A_E(y)$ does not entail $A(y)$: encoding is a relation different from the ordinary instantiation of properties by objects, typically expressed by the copula (hence the name of the approach). Quite generally, one can then have $A_E(\varepsilon x A(x))$.[1] Call this CP_E.

A third approach is now coming to be called *Modal Meinongianism* (MM): see Berto 2011, 2012, Priest 2005.[2] In a nutshell, it goes as follows. There are worlds other than the actual. Some are possible, and some are impossible.[3] According to MM, $A(\varepsilon x A(x))$ holds in full generality; but it may not hold at the actual world (though it may). All that can be guaranteed is that it holds in some world or other, namely those worlds that realize the situation envisaged by the person who uses the description. Call this version of Characterization CP_M. The name "Modal Meinongianism" is due to the fact that characterization is understood with reference to worlds other than the actual. Kroon's objections explicitly target MM.[4]

---

1. Actually, even in this case there is a restriction on $A(x)$, namely that it not contain mention of encoding, or paradox results.

2. We shall refer to the last of these, *Towards Non-Being*, as *TNB* in what follows. The Berto references argue that Modal Meinongiansm is preferable to the other approaches.

3. Possible worlds are now common currency among analytic philosophers. For an introduction to impossible worlds, see Berto 2009.

4. Berto 2012 formulates the four versions of characterization slightly differently, without mentioning descriptions at all. CP: For any condition $A(x)$, something satisfies $A(x)$. CP_N: For any nuclear condition $A(x)$, something satisfies $A(x)$. CP_E: For any condition (which does not

Another feature of MM is that it limits, of necessity, what Meinong called the Principle of Independence of *Sosein* (the having of properties by objects) from *Sein* (their existential status). *Some* properties are independent from existence, but some others (typically, those involving the having of causal features, or spatiotemporal location) are not: they entail it—at the actual world and, arguably, at all possible worlds. Kroon's objections to MM target such a limitation, as well as the CP_M.

Kroon has two objections. The first is to the effect that "MM is […] much more unfriendly to central Meinongian intuitions than its proponents allow" (24). This hinges on the fact that, in the MM theory, actually nonexistent objects cannot actually have existence-entailing properties. They can be characterized as having such properties, e.g., being a mountain and made of gold, but they can satisfy the characterization only at worlds different from the actual. Thus "in particular, it is *false*, not *true*, that the golden mountain is golden", and "such an outcome is bound to strike many contemporary Meinongians as a reason to reject CP_M" (27).

The second objection is to the effect that MM "cannot even guarantee that the golden mountain doesn't exist" (23), it "cannot even endorse the Meinongian truism that an object like the golden mountain lacks existence" (24). Given the way nonexistent objects are dealt with in MM, "nothing in Priest's theory allows him to conclude that the golden mountain lacks all [the existence-entailing properties], yet it must lack all of them for it not to exist" (28). Kroon rightly takes this second objection as the more serious.

2. *Preliminaries*

We will explain in detail the first objection, and take care of it, in Sections 3 and 4. We will do the same with the second objection in Sections 5 and 6. Before we set about these tasks, though, we need to clarify some issues concerning MM, and indeed Meinongianism in general. Kroon does not pay a lot of attention, we think, to some important distinctions; highlighting them is a useful preliminary for addressing his criticisms.

---

mention encoding), something encodes $A(x)$. CP_M: For any condition $A(x)$, something satisfies $A(x)$ at some world. One can then add different accounts of descriptions to this machinery, possibly with some extra conditions. Kroon, however, formulates the variants of the CP as in the text, and we follow him in this.

The first thing to notice is that, as a metaphysical theory, MM is not generally supposed to rule out the existence of things *a priori*. Did Homer ever exist? The city of Atlantis? That of Troy? It is preposterous to think that armchair metaphysics can address these issues all on its own: they are open to empirical investigation. We should not, then, expect MM itself to rule out that the golden mountain lacks existence. What Kroon therefore means in his second objection is that MM, *plus* what we know, largely *a posteriori*, about the world, cannot rule out the existence of things like the golden mountain.

But secondly, what kind of thing *is* the golden mountain—or, what does the description "the golden mountain" refer to? Kroon claims that MM "preserves Meinongianism's traditional commitment to nonexistent objects but offers a new account of their nature as objects and of the properties they might be said to have" (23). However, there is nothing like a single "nature" for nonexistents, whether one is a modal Meinongian or one of a more traditional kind. Meinongians are not committed, just because they claim that some things do not exist, to the nonexistence of some specific kinds of entity rather than others. One kind on which more or less all of them agree comprises purely fictional objects like Sherlock Holmes, Superman, Anna Karenina and Mr. Pickwick. But some treat mythical objects like Zeus or Thor as on a par with fictional objects, while others disagree. Some include mere *possibilia*—things that exist at other possible worlds but not at the actual one, like Wittgenstein's merely possible oldest daughter. Some Meinongians take abstract objects as nonexistent too, while others follow Meinong's original view and allow abstract objects to exist, though in a way different from concrete existents (often called "subsistence").[5] One of us (FB), following Routley (1980), takes seriously the view that also past objects like George Washington or Socrates are currently nonexistent objects. Besides, one of us (FB again) considers fictional objects, unlike *possibilia* and past existents, as necessarily nonexistent, while the other (GP) disagrees: for FB, there is no possible world where Sherlock Holmes exists, while for GP, there is. It is a mistake, then, to treat Meinongianism, even only of the modal kind, as if it was committed to a unique view on what does not exist. Meinongianism, and also Modal Meinongianism, can come in very different kinds.[6]

---

5. At least, this is so when the relevant abstract objects are consistent or well-defined: division by seven and the set of natural numbers exist/subsist, but division by zero and the Russell set do not. Thanks to Kroon for pointing this out.

6. *Quineanism* as well—as we may call the opposite view that everything exists—comes

The third issue we need to focus on is linguistic (semantic and pragmatic), and concerns the different ways in which referential expressions, in particular descriptions, acquire—when they do—a denotation. Meinongianism as such—as the claim that some things do not exist—does not commit one to any particular semantics and pragmatics of referential expressions, and specifically of descriptions. The admission of nonexistent objects *per se* does not even bring theoretical commitment to the claim that all singular terms denote (in all contexts of use). This is true of MM as of other forms of Meinongianism. Among Meinongians who subscribe to the nuclear version of the CP, Routley (1980) has all descriptions denote, whereas Parsons allows for non-referring ones, and claims that the issue is "primarily a linguistic question, or one of formulation" and does not entail "a serious ontological disagreement" (Parsons 1979, 653). Also, within MM, different accounts are possible—and actual: *TNB* goes for the view that all well-formed singular terms should denote, but Priest (2011a) and Berto (2012) explore versions with non-denoting terms.

With these provisos under our belt, we can start addressing Kroon's objections directly.

3. *Against literalism*

Kroon's first objection is to the effect that MM is contrary to "Meinongian intuitions" (24) or to the "spirit" of Meinongianism (27), or even plainly "strikingly counterintuitive" (24). Unlike naïve or nuclear Meinongianism, MM claims that characterized objects often don't really and literally have their characterizing properties at the actual world. In particular, nonexistent objects actually lack the existence-entailing properties they are characterized as having: "it is at the very least misleading for MM to claim that CP_M offers a viable sense in which the golden mountain is golden when it admits that no golden mountain is actually golden" (32).

Now, we demur from the thought that Meinongians perforce intuit that the golden mountain is literally golden and a mountain. The home of Meinongianism is the theory of intentionality. We can think of, and have other intentional relations to, objects, and some of these do not exist.

---

in extremely diverse forms. Some Quineans reject abstract objects, others admit them; some are presentists, other eternalists; some may count two or more objects in the same portion of spacetime, others always count at most one; some claim that *being* is spoken of in different ways, others that it is univocal, etc.

Thus we can think of the golden mountain. The object we think of had better be golden and a mountain in some sense—or what on earth are we thinking about? However, this does not have to be a literal sense (this does not, therefore, impact on the ability of the theory to "deal with the usual issues of interest to Meinongians", contra Kroon, 28). Indeed, it had better not be. Such a claim is not only false, it is opposed to common and much more robust intuitions. Let us see why, by focusing on what all Meinongians take as the most uncontroversial kind of non-existents: purely fictional objects like Sherlock Holmes.

The alleged "intuition", shared by naïve and nuclear Meinongianism but not by MM, consists, in fact, in the former views' being affected by what Kit Fine (1982) called *literalism*: the idea that non-existents like Holmes literally and really have the (nuclear) properties they are characterized as having (in the relevant fictions). Clearly, nuclear features, like those of being a detective and of living in 221b Baker Street, are ascribed to Holmes in the Doyle stories. Naïve and nuclear Meinongians want these properties to be had by Holmes at the actual world. MM denies this, as Kroon rightly stresses (26f.): for modal Meinongians like us, Holmes is a detective and lives in 221b Baker St., only at the worlds that realize the characterization provided by Doyle, not in actuality.

Now, we ask literalists: how could Holmes literally possess those features? In reality, Baker Street 221b hosted an enterprise, the Abbey Road Building Society, and it has never been the house of any private detective. It is literally *false*, not true, that 221b is, or has ever been, Holmes' home. In one of the Doyle stories we are told that Holmes has tea with William Gladstone (the example is due to Woods 1974). How can this be literally true? William Gladstone is a real (past) existent, who certainly never had tea with any purely fictional object.

One may claim that Holmes actually lived in a nonexistent 221b Baker Street, or had tea with a nonexistent doppelganger of Gladstone. But this multiplication of objects is itself counterintuitive. Fictional stories include lots of references to nonfictional objects, which are only represented in the stories as interacting with purely fictional ones. Napoleon features in *War and Peace*, and Napoleon was a very existent man. MM is not forced to treat "Napoleon" as ambiguous, as it happens in forms of realist abstractionism about fictional objects à la van Inwagen (1977, 51): (a) normally denoting the historical character, i.e., the concrete and (formerly) real man; but also (b) referring to a quite different abstract object, when the name occurs in extra-fictional discourse on the literary character of *War*

and Peace; and, perhaps, also (c) denoting nothing at all, when it occurs in the intra-fictional discourse of *War and Peace*. Such ambiguities seem to be introduced *ad hoc*, because they are not confirmed by the intuitive data: competent speakers have no sense of the postulated ambiguity. As the Wikipedia entry on *War and Peace* claims: "There are approximately 160 real persons named or referred to in *War and Peace*".[7]

Besides implying claims that are in point of fact false, literalism severs intuitive nexuses between properties, and specifically between various properties and existence—which, to Meinongians, is but yet another non-blanket feature of individuals. As a nonexistent, Holmes cannot literally have features that entail existence, like living in a real street, having tea with Gladstone, or being a detective. If something is a detective and lives in a London street, then it is natural to think that it is a human being, a physical object, a spatiotemporal occupier, and endowed with causal properties. Asking where the person is, or why, as a detective, they cannot help the metropolitan police to solve crimes, is quite sensible: things lacking real existence cannot really have existence-entailing properties involving causal features or spatiotemporal location.[8]

The "Meinongian intuitions" discharged by MM are precisely literalist intuitions. We claim that they *have* to be disrespected, because they are just wrong. Nor is MM the only form of non-literalist Meinongianism on the market. Dual copula Meinongianism also denies that Holmes literally is a detective living in Baker Street. As a nonexistent object, for dual copula

7. See http://en.wikipedia.org/wiki/War_and_Peace. As Kroon pointed out to us, some combinations of descriptive expressions and names, like "the Napoleon of *War and Peace*", may still provide some support to the ambiguity view: cf. "The Napoleon of *War and Peace* is a cleverly constructed character, very different from the real Napoleon". Berto (2011) proposes that even these fail to force a multiplication of referred objects: both "the Napoleon of *War and Peace*" and "the real Napoleon" refer to the one Napoleon. The price to be paid is that one needs to paraphrase something away: the former expression should be read as something like "Napoleon, as represented in *War and Peace*", and the latter, as "Napoleon, as he really is/was".

8. Here's how the point is nicely made by Nathan Salmon: "Undoubtedly, existence is a prerequisite for a very wide range of ordinary properties—being blue in colour, having such-and-such mass, writing *Waverley*. But the sweeping doctrine that existence universally precedes suchness has very clear counterexamples in which an object from one circumstance has properties in another circumstance in virtue of the properties it has in the original circumstance. Socrates does not exist in my present circumstance, yet he has numerous properties here—for example, being mentioned and discussed by me. Walter Scott, who no longer exists, currently has the property of having written *Waverley*. He did exist when he had the property of writing *Waverley*, of course, but as every author knows, the property of writing something is very different from the property of having written it. Among their differences is the fact that the former requires existence" (Salmon 1998, 290f.).

Meinongians Holmes only encodes these properties, encoding being distinct from exemplification. Holmes can exemplify and literally have lots of properties, such as being nonexistent, being self-identical, being thought of by us, etc. But the features ascribed to him in the Doyle stories are for the most part only encoded, not literally exemplified, by Holmes. So MM's non-literalism does not even make it an isolated faction among the Meinongian tribes.

And not only are those literalist "Meinongian intuitions" wrong: they are not even *common-sense* intuitions. We agree with the arguments of Sainsbury (2010, 26ff.), to the effect that people don't even share the belief that Holmes really is or was a detective, and really lives or lived in 221b Baker Street. People do not believe such claims as "Holmes lived in 221b Baker Street" to be correct descriptions of actuality. It is generally agreed that, if someone believed something like this, she would stand in need of being corrected by those who know better.

Think of a London policeman replying to a tourist asking where 221b Baker Street, the famous residence of Sherlock Holmes, is located: "Sir, Sherlock Holmes does not exist and has never existed; it's just a fictional character due to the novelist Conan Doyle. Baker Street, well, that does exist: it's just down there. But Holmes didn't really live there: he only lived there according to Doyle's stories".

The policeman gets it right, in the non-literalist way, by just relying on common sense. Intra-fictional ascriptions of existence-entailing properties like being a detective or living in Baker Street are to be understood as implicitly prefixed by an "according to the story", non-factive clause. This is often omitted in conversation but contextually easily understood (as noted in *TNB*, 117, fn. 2). We move seamlessly from truth in reality to truth according to a fiction and back all the time. An historian lecturing on the ancient Greeks' religion claims: "Zeus is the king of the Greek pantheon, living on Mount Olympus, …" etc. We understand him as speaking the truth, for we know he means: that was so according to the Greek mythology, not in reality.

4. *What we do all the time with "a/the F"*

Literalism is thus both false and unintuitive, and MM is right in rejecting it. A linguistic point is left open, though, and clearly spotted by Kroon in Section 3 of his paper. MM's non-literalism entails that, unlike what

happens with literalist (naïve and nuclear) Meinongianism, many definite and indefinite descriptions will have to refer to things that don't currently or actually satisfy the relevant conditions. Kroon's example, mentioned above, involves "the golden mountain" referring to something that is neither a mountain nor golden at the actual world. But we can equally well keep using Holmes as our chief example. According to MM, nonexistent Holmes is not really a detective, nor does he really live in Baker Street. However, we use such features to build descriptions apparently successfully referring to him. We felicitously refer to Holmes as "Doyle's detective living in 221b Baker Street". How come?

We reply that even this is not a theoretical minus of MM, because we do felicitously use descriptions to refer to things not actually or currently satisfying them all the time; nor is this an issue having specifically to do with Meinongianism, or nonexistence. Let us see why.

Donnellan's (1966) famous referential/attributive distinction might well be taken to show that we can use descriptions to refer to objects that don't actually or currently satisfy them. "The man over there with the champagne in his glass is happy" successfully refers to a man in the corner who is, as a matter of fact, happily drinking sparkling water. Kripke's (1977) rejoinder to Donnellan is also well-known. We should distinguish between *speaker referent*, the object a speaker intends to refer to, and *semantic referent*, what is literally referred to. Only the latter has to do with semantics properly, whereas investigation of the former falls in the realm of pragmatics. Our intention to refer to a person who, unbeknownst to us, has water in his glass, does not affect the proposition literally expressed by our utterance of "The man over there with the champagne in his glass is happy": this does not depend on the speaker's intentions but on the description's semantic denotation, which cannot be any non-champagne drinker.

But even if one accepts the distinction between speaker referent and semantic referent, it remains the case, as all may agree, that the semantic referent of a description is context dependent. Thus, in "The President wants to see you", "the President" will normally refer to different people if this is said in the White House or the Bundestag. And the person picked out in a context may not be the person who *actually* satisfies the condition. This can happen because of spatial displacement. Thus, suppose that we are in the USA, where the current president is Barack Obama. Yet we talk about Germany, and you say "The President may be the head of state, but actually, the person who runs the country is the Chancellor, Angela Merkel": "the President" would then refer to Joachim Gauk. Or, it can

happen because of temporal displacement. Thus, suppose that we are in the UK, and the monarch is Elizabeth II. However, we are discussing the life of Shakespeare, and you say that the Bard never met the Monarch: "the Monarch" would refer to Elizabeth I.

As so often happens, there is a modal analogue of this temporal phenomenon. The Earth is the third planet from the sun. However, suppose we are discussing an envisaged situation where the Solar System is pretty much as it actually is, but there is a sub-Mercurial planet, Vulcan. In this context, "the third planet from the sun" refers to Venus. Which brings, us, of course, very close to fictional objects. The smartest cocaine-using detective in London is probably some corrupt member of the Metropolitan Police Force. But if we are talking about the world as described by Conan Doyle, "the smartest cocaine-using detective in London" certainly refers to Holmes. Definite descriptions, then, can semantically refer to things that do not actually have the features in question. And this can be the case whether or not the object in question exists—as the previous examples make clear.

Semantic reference is, as we have noted, context-dependent. And what context we are in depends on many factors, one amongst which concerns the intentions of the speaker (does one mean to be talking about an historical epoch? Or about a hypothetical scenario? Etc.). So even for semantic reference, it is worth noting, intentions do get in the act. Incidentally, this is explicitly acknowledged in *TNB* (where choice functions are integral to the denotation of a description, as we shall see in detail in due course):

> The deployment of a choice function is a recognition of the fact that, as far as the formal semantics go, the denotation of the descriptive term is non-deterministic. That is, the denotation of the term is something that is determined by factors outside the semantics. Principal among these is context, and especially speaker intention. […] Thus, suppose you say (truly), for example: 'I saw a man on the tram I was on yesterday; he looked rather sad.' The referent of 'a man on the tram I was on' in this context is the particular man whom you saw, and to whom you now intend to refer. (Note that there could have been more than one sad-looking man on the tram; but you are talking about a particular one of them.) Of course, you could be lying: the man on the tram was not sad. The description refers to him none the less. Maybe you didn't even get on a tram at all. In that case, the description refers to the presumably non-existent object intended in your imagination. (*TNB*, 94)

5. De re *reference fixing*

We now turn to Kroon's second objection. He claims that MM, together with what we know about the world, "cannot even guarantee that the golden mountain doesn't exist". After stressing that characterized objects in MM may often lack the properties they are characterized as having at the actual world, he claims that what would be even "far less acceptable" (27) is the golden mountain's having existence-entailing properties at the actual world:

> Whatever is it like at the other worlds, at the actual world the golden mountain is not granitic or silver, nor is it located on Uranus, or in any other part of the universe, for it does not exist at the actual world, and to have any existence-entailing property at a world a thing has to exist at that world. And the reason we know it does not exist at the actual world is that nothing at the actual world is uniquely a golden mountain. (28)

However, Kroon continues, MM cannot accept this conclusion. Even once we know that nothing is a golden mountain at the actual world, MM cannot sustain the right reply to the question: "How do we know that nonexistence is among its [the golden mountain's] properties?" (28). We cannot say that the object characterized as *a/the A* "possesses no existence-entailing properties, *even* when we know that there are no *A*s" (29). The theory licenses only the claim that the golden mountain exists at the worlds (or at least, at the possible ones) where it has the properties it is characterized as having, namely where it is a mountain and made of gold; but the theory is silent on the actual existential status of the object.

Which object? As we argued in the previous section, the referent of a description is usually context (and intention) relative. Then there is mostly no unique answer to the question. While Kroon apparently acknowledges this (as we will see), he develops most of his objection plainly talking of the golden mountain as if what was referred to by the description was, context-independently, a unique thing. This is not so, though. MM cannot give a *single* reply to Kroon's question, "How do we know that nonexistence is among the golden mountain's properties?" "The golden mountain" can refer to things of quite different kinds in different contexts, and quite different replies to the question will have to be given in such different contexts. Sometimes we will know that the object referred to by the description does not exist on the basis of our empirical information about the actual world. Sometimes we will know that it does not exist on

the basis of our knowing the kind of thing at issue (and, of our having the right ontological account of things of that kind). Sometimes we will even know that the thing at issue is *existent*—and also, if such be the case, that it is grey and granitic. Context and speaker intentions will make all the difference.

Here's one context where the description refers to an existent object. We often make up stories by intentionally referring to real, existent objects, which we nevertheless characterize via properties they actually lack: we use them as props in games of make-believe. Kroon acknowledges that "sometimes our imaginative activities are directed at existent things" (29). Children can pretend that the elm in the garden is a magic tree, or that their bike is a Harley 883. Similarly, we can start to tell the following story:

> Imagine that when Edmund Hillary first climbed mount Everest he discovered that, because of some peculiar geological phenomenon, its summit was largely made of almost pure gold. Soon expeditions were organized from different countries to reach the top of the mountain: everybody wanted to take advantage of the golden mountain and many international controversies began …

In the context created by our story, "the golden mountain" obviously refers to mount Everest. We single out an existent by telling a story *de re* about it. Then we refer to it as the golden mountain, and it is understood that the thing is only represented as being such within the story, without it actually being so.

In such a context, it is not true that "whatever it is like at the other worlds, at the actual world the golden mountain is not granitic or silver" (Kroon, 28) because it doesn't exist. We know what "the golden mountain" denotes in such a context, and we know that Everest is actually existing, grey and granitic.[9] This very existent object has the property of being golden at the worlds that realize our story, those of being grey and granitic at the actual world, and that of being a mountain both at those worlds and in actuality.

9. We may take the Everest as grey and granitic for the sake of the argument. Having such properties is arguably a matter of degree for mountains. As it happens, anyway, mount Everest does include substantive amounts of grey stones and granite (see http://en.wikipedia.org/wiki/Mount_Everest).

6. *Where Kroon is right*

In the sort of context last envisaged, one takes an existent object, refers to that *de re*, and then imagines a non-veridical situation about it. There is a quite different sort of context, however. And about this, Kroon has a very valid point to make. One can hypothesize or imagine a certain scenario, and then one can refer to an object in that scenario. Thus, one might postulate the existence of a sub-Mercurial planet, or imagine and start to write down a story about an eccentric detective. In this sort of situation, the reference of the description is not parasitic on some prior act of reference-fixing. In such cases, the CP_M itself determines how the reference of the description is fixed.

TNB, 92-3, formulates the CP_M as follows (simplifying slightly to avoid irrelevant complications, and where @ is the actual world):[10]

(i) If something satisfies $A(x)$ at @, $\varepsilon x A(x)$ denotes one such thing.
(ii) If not, it picks out some object or other which satisfies $A(x)$ in the situation one is envisaging.

Formally, the picking out in each case is done by a suitable choice function; informally, this represents an intentional act (the intentional act can be construed in both a realist and a non-realist way: see Priest 2011a). Now there is nothing in this account which requires $\varepsilon x A(x)$ not to exist. Nor should there be, at least as far as clause (i) is concerned. Let $A(x)$ be "x is a sub-Mercurial planet responsible for the precession of Mercury's perihelion". Then, as a matter of fact, nothing existent satisfies $A(x)$. But had the world been different, the description could have referred to an existent object: had $A(x)$ been satisfied at @, the description would have denoted a planet, therefore a causally efficient object, therefore (for MM) an existent.

However, there is an issue with clause (ii). Nothing in this case requires $\varepsilon x A(x)$ to refer to a non-existent object either (as *TNB*, 92, does point

10. Kroon, fn. 7, takes *TNB* to task for calling descriptions rigid, suggesting that what it should say is that the expression "the object represented as being the golden mountain" is rigid. But *TNB* means exactly what it says. Once the denotation of $\varepsilon x A(x)$ is picked out, the description refers to that object in every world (which is not, of course, to say that the object satisfies $A(x)$ in every world). Besides, a non-rigid semantics for descriptions is also possible, as explained in *TNB*, 93. Kroon also says, fn. 9, that for MM, if there is an actuality operator in the language, this must work differently at possible and impossible worlds. He cites Beall as showing that "this is a serious weakness". It is not, as is shown in Priest 2011b, 3.3.

out). But this seems wrong. In such cases the term should refer to a non-existent object. Thus, Vulcan does not exist; neither does Sherlock Holmes; neither does Zeus (where these names are to be taken as abbreviations for an appropriate description). Here Kroon is exactly right.

The change to the theory to rectify the matter is, however, very simple. Clause (ii) can be reformulated as:

(iii) If not, it picks out some *non-existent* object or other which satisfies $A(x)$ in the situation one is envisaging.

(Here the non-existence is actual: the object may well exist in the relevant non-actual situations, of course). With this change, if nothing satisfies the characterizing condition at @, the object referred to does not exist. So none of Vulcan, Sherlock Holmes, and Zeus, exists.

We can, in fact, pack clauses (i) and (iii) into one, showing how there is a uniform act of intentionality. Let α abbreviate "At @, some y is such that $A(y)$". If Φ is a choice function on sets of objects, then the denotation of $\varepsilon x A(x)$ is:

(iv) $\Phi(\{x : (\alpha$ and $A(x))$ or $(\neg\alpha$, x is a non-existent object, and the envisaged world is one where $A(x)\})$.

One worry one might have here is that this is an *ad hoc* modification of the theory. But it is not: it simply rectifies an oversight in the original formulation of *TNB*. When we construct a theory of intentionality with its denizens of objects, existent and non-existent, we are trying to account for the obvious data: that one can think of things whether or not they exist, that we can tell a story about mount Everest in which it is golden, and that Vulcan does not exist. All theorising, including our modification, is *ad hoc* in this unobjectionable sense. The new version of the theory simply takes into account a bit of data that had been overlooked. Such *ad hocness,* thus, is quite unlike the one mentioned in Section 3, where an ambiguity which is not supported by the data is needed to defend a literalist view.

One should note that, in the last instance, any theory of descriptions is constrained in such an *ad hoc* way. To illustrate: all can agree that if something satisfies $A(x)$, '$\varepsilon x A(x)$' refers to such a thing. The problem is what to do in the other case. If one is not a noneist, then, in such a situation, we have a case of reference failure. How then to proceed? One possibility (Frege's) is to assign the description an arbitrary denotation—say the

empty set. Let us take 'Sherlock Holmes' to be short for an appropriate description. Then this policy will make 'Sherlock Holmes exists' true. And what is wrong with that? Simply that it gets the data wrong.

Another policy is to make every atomic sentence false by definition (say the contextual definition of Russell). Alternatively, however, we might make every such sentence true by definition. And what is wrong with that? Again, it gets the data wrong. This policy makes 'Sherlock Holmes lives in Beijing' true. Any policy concerning descriptions must be constructed to do justice to the data in such a way. That is what the theory is *for*.

Coming back to the present proposal, another worry one might have concerns how one manages to intend a non-existent object. Kroon himself raises this worry:

> [T]he response depends on an author's having the ability to intend a non-existent object, knowing *a priori* that the object she thereby selects is indeed nonexistent. It is difficult, however, to make sense of such an ability. How can the agent know *a priori* that the object she manages to select is *in fact* nonexistent? […] Couldn't the agent intend what she takes to be a nonexistent object, but just make a mistake? We can make mistakes when intending an existent object; we might be hallucinating "that mountain in the distance", for example. So why not when intending a nonexistent object? (32)

Now, first, on the above account, one cannot know *a priori* that the object intended does not exist, since one cannot know *a priori* that nothing at @ satisfies $A(x)$ (at least for possible conditions). As we made clear at the beginning of this paper, this cannot in general be settled by the MM theory as such: it depends on how things turn out in the world. In particular, one might intend the description to refer to an existent object (as did the scientists who postulated Vulcan), but the intention may not be realised.[11]

If, however, nothing actually satisfies $A(x)$, the denotation of the description, the object intended, is a non-existent object. One may hear Kroon as asking: How so? Indeed, how does one intend an object with any properties, particularly the one defining the set on which the choice function in (iv) operates? One answer is that it is the very nature of intentionality to single out an object of a certain kind, and given a bunch of objects, one can just point mentally to one, in the same way that, given a bunch of

11. The distinction between intending to (aiming to) refer to a non-existent object, and the object actually intended (as the target of the act of reference) being non-existent, may clear up some misunderstanding in the personal communication referred to by Kroon (31).

physical objects, one can point physically to one of them[12] (indeed, an act of physical pointing presupposes the intentional act that goes along with it: otherwise, one could be pointing at many things). And if one imagines an object with certain properties then, *ex hypothesi*, what one does *is* imagine such a thing. This is not a defense of psychological infallibilism. We may, indeed, think that our mental state is something that it is not. Rather, it is the phenomenological analogue of Kripke's (1971) point that a possible world where Humphrey won the election is, *ex hypothesi*, a world where *Humphrey* won the election—and, we might add: *won the election*. At any rate, this is a quite general issue with MM, and the revised denotation conditions for descriptions do nothing to make the matter better or worse.[13]

Let us finally, in this section, see how the revised definition addresses Kroon's Hillary counter-example (32-3). This is as follows. At 8,848m above sea level, Everest is the tallest mountain in the world, and Hilary climbed it once. Suppose that Fred believes all this—except that he thinks that Everest is 10,000m above sea level. He then imagines the highest sub-9,000m mountain in the world which was climbed by Hillary twice—call it "H2". He is confident enough that this does not exist, but is not certain, so does not intend (aim) to refer to a non-existent object. Now, H2 presumably has ordinary modal properties, including the property that there is a world w in which H2 is the highest sub-9,000m mountain and was climbed by Hillary just once. The question for Kroon is how MM can rule out the actual world being such a w: if it is, then H2 is Everest and H2 does exist after all.

The answer is that it is the non-existence of H2 that rules this out. There indeed are worlds w of the kind described, but the actual world is not one of them. For let $A(x)$ be the condition "x is the highest sub-9000m mountain in the world and Hillary climbed x twice". According to the account just given, since nothing satisfies this condition at the actual world, $\varepsilon x A(x)$, that is, H2, does not exist. So while Fred did not intend to refer to a non-existent object, the object intended, hit by the act, was non-existent nonetheless.

---

12. This answer is defended in Priest 2011b, 1.5. Some people find it more plausible that this can happen in the case of non-existents if the intentional act creates the object, in a certain sense: see Priest 2011c, and Berto 2012, Chapter 9. Of course, "create" here cannot mean "bring into existence". Rather, it means to extend the domain of discourse with a new object.

13. Though it certainly increases the theory's "intentional-metaphysical load", as Kroon put it in correspondence.

7. *Conclusion*

In the Preface to *Towards Non-Being*, Priest claimed:

> Nor do I take the version of the [MM] view presented here to be definitive. A number of the techniques developed in the book are relatively novel and untried, and I would be surprised, indeed, if better techniques could not sometimes be found. (Priest 2005, *x*)

Indeed so. Most of the points raised by Kroon in his paper can be addressed by MM—and some of them quite easily, as we have seen. But Kroon's final point forces us to declare the initial MM account, as presented in *TNB*, in need of revision. We will still be surprised if no further revisions turn out to be required in the light of future inspection. By triggering the one just described, Kroon's paper is, we think, the most perceptive criticism of MM to date.[14]

References

Berto, Francesco 2009: "Impossible Worlds". *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/impossible-worlds/.

— 2011: "Modal Meinongianism and Fiction: the Best of Three Worlds". *Philosophical Studies* 152, 313–335.

— 2012: *Existence as a Real Property*. Dordrecht: Synthèse Library, Springer.

Donnellan, Keith 1966: "Reference and Definite Descriptions". *Philosophical Review* 77, 281–304.

Fine, Kit 1982: "The Problem of Non-Existents." *Topoi* 1, 97–140.

Kripke, Saul 1971: "Identity and Necessity". In: Milton K. Munitz (ed.), *Identity and Individuation*. New York, NY: New York University Press, 135–164.

— 1977: "Speaker Reference and Semantic Reference". In: Peter A. French, Theodore E. Uehling Jr & Howard K. Wettstein (eds.), *Contemporary Perspectives in the Philosophy of Language*. Minneapolis: University of Minnesota Press, 6–27.

Kroon Fred 2012: "Characterization and Existence in Modal Meinongianism". *Grazer Philosophische Studien* 86, 23–34.

Parsons, Terence 1980: *Non-Existent Objects*. New Haven, CT: Yale University Press.

14. We are grateful to Fred Kroon and to an anonymous referee of this journal for their helpful comments on the paper.

Priest, Graham 2005: *Towards Non-Being*. Oxford: Oxford University Press.
— 2008: "The Closing of the Mind: How the Particular Quantifier Became Existentially Loaded Behind our Backs". *Review of Symbolic Logic* 1, 42–55.
— 2011a: "Creating Non-Existents". In: Frank Lihoreau (ed.), *Truth in Fiction*. Heusenstamm: Ontos, 107–118.
— 2011b: "Against Against Non-Being". *Review of Symbolic Logic* 4, 237–253.
Rapaport, William 1978: "Meinongian Theories and a Russellian Paradox". *Noûs* 12, 153–180.
Routley, Richard 1980: *Exploring Meinong's Jungle and Beyond*. Canberra: RSSS, Australian National University.
Sainsbury, Mark 2010: *Fiction and Fictionalism*. London & New York: Routledge.
Salmon, Nathan 1998: "Nonexistence". *Noûs* 32, 277–319.
Van Inwagen, Peter 1977: "Creatures of Fiction". *American Philosophical Quarterly* 14, 299–308.
Woods, John 1974: *The Logic of Fiction: a Philosophical Sounding of Deviant Logic*. The Hague: Mouton.
Zalta, Edward 1983: *Abstract Objects: an Introduction to Axiomatic Metaphysics*. Dordrecht: Reidel.
— 1988: *Intensional Logic and the Metaphysics of Intentionality*. Cambridge, Mass.: MIT Press.

Grazer Philosophische Studien
90 (2014), 201–212.

STANDARD GETTIER CASES: A PROBLEM FOR GRECO?

Shane RYAN
Soochow University (Taiwan)

Summary
I argue that Greco's handling of barn-façade cases is unsatisfactory as it is at odds with his treatment of standard Gettier cases. I contend that this is so as there is no salient feature of either type of case such that that feature provides a ground to grant, as Greco argues, that there is an exercising of ability in one type of case, standard Gettier cases, but not in the other, barn-façade cases. The result, I argue, is that either Greco must revise his grounds for treating barn-façade cases as he does or he must revise his treatment of standard Gettier cases.

1. *Greco's account of knowledge*

On Greco's account of knowledge, S knows that p if, and only if, S believes truly *because* of S's cognitive ability.[1] For example, S's believing the truth that 28 x 9 is 252 is explained by S's believing from a cognitive ability, specifically a mathematical ability. For Greco, it's in the nature of cognitive ability, and ability generally, to be a reliable process grounded in the cognitive character of an agent.

In Greco's account, no separate condition is used for dealing with Gettier-type cases (Greco 2010, 75). This is an attractive feature, as not requiring such a separate condition makes for a theoretically elegant account. No separate condition is required as Gettier cases are ruled out as cases of knowledge because they don't meet the criterion of being a true belief *because* of cognitive ability.[2]

That Greco claims that knowledge is *exclusively* accounted for as a true belief because of "epistemically virtuous belief forming processes",

1. Greco (2009, 318) sees intellectual abilities as a species of intellectual virtue.

2. This section is intended to provide a basic overview of Greco's account of knowledge. In sections 3 and 4 more is said about his position, particularly with regard to specific types of Gettier cases.

means that the kind of virtue epistemology which he endorses is what Pritchard (2010, 24) has termed robust virtue epistemology. Weak virtue epistemology is an alternative to robust virtue epistemology, of which Pritchard (2010) argues in favour. According to weak virtue epistemology a condition separate from a virtue condition is required to rule out Gettier cases.

2. *The Barney Case*

First, let us turn to the Barney case (also known as the Barn-Façade case) as described by Pritchard (2012, 251):[3]

> Using his reliable perceptual faculties, Barney non-inferentially forms a true belief that the object in front of him is a barn. Barney is indeed looking at a barn. Unbeknownst to Barney, however, he is in an epistemically unfriendly environment when it comes to making observations of this sort, since most objects that look like barns in these parts are in fact barn façades.

What's of note in this case is that, unlike in standard Gettier-type cases, there appears to be a true belief because of the exercise of cognitive ability.[4] There is no so-called lucky intervention involved which results in the agent's belief turning out to be true. Rather, luck is said to be present in that had the agent come to believe on the same basis in modally close cases, then the agent could have easily gained a false belief.[5] So had Barney formed his belief on the same basis in closeby possible worlds, for example, ones in which his eyes happen to fall upon the barn façades in the county rather than the few barns, then he would form false beliefs.

Pritchard (2010, 35f.) argues that Greco's robust virtue epistemology problematically rules in Barney-type cases as being cases of knowledge. More specifically, he argues that Greco's criterion for knowledge in Barney-type cases is met, there is true belief because of cognitive ability. The standard intuition among epistemologists, however, is that protagonists

3. The original barn façade case first appeared in a paper by Alvin Goldman (1976). Goldman credits the example to Carl Ginet.

4. It should be noted that I am here representing Greco's account as being one on which an exercise of ability rather than ability is necessary for knowledge. It seems perfectly plausible that there may be occasions on which an agent does have the relevant ability which, if exercised, has the potential to gain the agent knowledge, but which, on those occasions, is not in fact exercised.

5. Pritchard (2010, 36) describes Barney-type cases as involving environmental epistemic luck.

in Barney-type cases don't know, rather they have justified but lucky true believes.

3. *Greco's handling of the Barney Case*

An early response from Greco (2009b, 21) to the Barney case, and implicitly to cases of that type generally, is to argue that no such environment-relative cognitive ability is exercised.[6] For Greco (2010, 10) an environment-relative ability to Φ implies a reliable environment-relative ability to Φ. Cognitive ability plays the same kind of role as justification does in more traditional epistemic accounts, so incommodiously for Greco, it seems that he might have to say, if using more traditional epistemological language, that Barney-type cases are not cases of justified true belief but rather are cases of mere true belief. Certainly, they are cases, on Greco's view, in which there is not knowledge conducive justification; the relevant environmental ability, the exercise of which is necessary for knowledge, is not present.

But what does Greco mean by an environment-relative ability? Greco and Turri (2011) write that, on the former's account, abilities are always environment-relative. We can't exercise our abilities in just any conditions. Greco gives the example of Tiger Woods's ability to sink putts being conditional on the environment being a certain way. We wouldn't say he has the ability to sink putts in hurricane-force winds. Elsewhere Greco (2010) claims that abilities are condition-relative as well as environment-relative. For example, whether a baseball player has the ability to bat well will depend not just on things like the weather, but also conditions such as him not having sand in his eyes.[7]

Greco (2010, 76f.) foresees a species of the generality problem hampering assessment of whether there is an exercise of ability in the Barney case. If we ask whether Barney has an ability relative to an environment, with the environment fixed at one particular level of generality, that of

6. It seems fair to say that this is still the general thrust of his response, but, as we shall see, he argues, with regard to the Barney case, that further description of the case is required.

7. It seems more appropriate just to talk about conditions; after all, surely environmental factors relevant to whether an ability can be exercised are simply conditions of a kind. Greco (2010, 77) writes that the two may overlap but that for his purposes "environment" should be thought of as "sets of relatively stable circumstances and 'conditions' as sets of shifting circumstances within an environment."

Barn Façade County, then the answer is no. If we ask whether Barney has an ability relative to different levels of generality, for example the particular farm upon which one real barn stands and no barn façades stand or across the globe upon which the number of real barns vastly outnumber the number of barn façades, then the answer is yes. This consideration raises the question of how we are to fix the appropriate level of generality of the environment, which is necessary to do if we are to judge whether an environment-relative ability is present in a particular case.

Greco (2010, 78) claims that the level of generality of an environment, when determining whether or not an ability is present, should be fixed "according to the interests and purposes of relevant practical reasoning." By way of illustration, Greco offers an example of how the level of generality can be fixed by a "practical reasoning context". Suppose S says that R "has the ability to hit baseballs"; what is being claimed about R's ability will differ depending on whether S is a baseball executive discussing whether to trade R, or S is a Little League coach discussing what role to give a new seven year-old player; the conditions and environments that the abilities of each player are relative to can be expected to differ enormously (Greco 2010, 78). The Barney case, Greco (2010, 79) writes, is underdescribed in that it is not clear what practical interests are in play, and, as discussed, we need to know that in order to be in a position to determine the generality of the environment to which Barney may or may not have a relative ability.[8] So without knowing this we are not in a position to judge whether Barney is exercising the relevant environment-relative ability, and ultimately whether he knows.[9] Greco (2010, 80) does offer a further description of the Barney case in which he adds that Barney is in the area to calculate property taxes, and that barns and barn façades are liable to different rates of tax. Greco's analysis is that Barney does not possess the

8. Greco, drawing on Edward Craig's (1990) *Knowledge in the State of Nature*, also claims that "the concept of knowledge is used to flag good information and sources of information for use in practical reasoning." (Greco 2010, 78). It's not clear to me, but it looks like he is saying that knowledge conforms to the same sort of pattern as ability does, in that whether someone has a cognitive ability that is such that knowledge may be gained because of its exercising will depend on practical considerations. And when it comes to cognitive abilities, specifically in relation to knowledge, we can identify what some of those considerations are; for example, flagging good information and sources of information.

9. Of course having an environment-relative ability does not imply the exercising of that ability, although the exercising of an environment-relative ability does imply the possession of that ability.

relevant environment-relative ability, and therefore there is no exercise of the relevant cognitive ability.

4. *Greco's handling of the Roddy Case*

What is crucial on Greco's (2010, 101) account of knowledge is that a belief is true because of ability rather than there simply being a coincidence of the exercise of ability and true belief. So there must be an environment-relative ability that is exercised and the exercising of that ability must explain why the belief that is gained is true; an agent happening to have a true belief and the relevant environment-relative ability exercised is not enough.

To see a motivation for Greco's claim that knowledge is true belief *because* of cognitive ability, consider the Roddy case from Chisholm (1977, 105), which has been adapted by Pritchard (2012, 251):

> Using his reliable perceptual faculties, Roddy non-inferentially forms a true belief that there is a sheep in the field before him. Unbeknownst to Roddy, however, the truth of his belief is completely unconnected to the manner in which he acquired this belief since the object he is looking at in the field is not a sheep at all, but rather a sheep-shaped object which is obscuring from view the real sheep hidden behind.

Intuitively Roddy doesn't know. Although a justified true belief is present, or, in Greco's terminology, a true belief and the exercise of cognitive ability are present, the belief's being true is not explained by the subject's present exercise of cognitive ability and therefore it is not a case of knowledge. Roddy seems not to know precisely because the truth of his belief lacks the appropriate causal connection to the exercise of his cognitive ability. It's a theoretical plus for Greco's (2009b, 19) account that its diagnosis of why Roddy doesn't know appears exactly right.[10] Furthermore, this is how Greco (2009b, 19–21) purports to handle standard Gettier-type cases generally.

---

10. The later Greco would presumably want to know what practical interests are in play in the Roddy case in order to determine whether Roddy is indeed exercising an environment-relative ability. There seems something costly about always having to identify practical interests in order to be in a position to determine whether an agent is justified in their belief, but I won't explore this thought any further here.

5. *The Barney and the Roddy Cases taken together*

As we have seen, Greco's treatment of Barney-type cases differs from his treatment of the Roddy case and standard Gettier-type cases generally. The treatment of the two types of cases differs in that Greco claims there to be an exercise of cognitive ability in standard Gettier-type cases and not in Barney-type cases. In his treatment of the Roddy case and standard Gettier-type cases generally, one of the most appealing aspects of his robust virtue epistemic account is on display. There seems to be no principled reason, however, for Greco to claim that there is an exercise of cognitive ability in the Roddy case, but not in Barney-type cases. To put the issue into sharper relief, why think that Roddy has an environment-relative ability in his case, according to which there is a sheep shaped object in his vicinity, but that Barney doesn't have an environment-relative ability in his case, according to which there are Barn façades in his vicinity?

To see what's at issue I consider two cases intended to be analogous to the Roddy case and Barney case respectively. First, however, some background details of the two cases need to be spelt out.

Tony lives in part of the Amazon Rainforest. Together with Tony live hundreds of x-type birds. Sometimes Tony goes up to a tree, listens carefully and identifies the presence of an x-type bird in the tree by its singing. In fact when there is an x-type bird in the tree and it is singing and Tony is at the foot of the tree listening carefully, Tony can reliably identify an x-type bird as being in the tree. Let's say that Tony is his tribe's tax inspector. "Householders" are liable to varying rates of tax depending on the number of trees on their properties that have at least some x-type birds in them when Tony is doing his rounds.[11]

For the first time ever some y-type birds, a cousin species of x-type birds, have ended up in Tony's part of the rainforest. More precisely, five y-type birds have gotten lost during the annual migration of y-type birds and the five spend a day in Tony's part of the rainforest before moving on.

Lost Birds, case one:
On one occasion that day, Tony ventures up to the foot of a tree, listens carefully, hears singing, and forms the true belief that there is an x-type

11. I realise the romantic image of a man in the woods taking time to listen to birdsongs has been ruined by allowing talk of property and taxes to intrude, but I do so in order to meet Greco's requirement of describing the practical interests in play and this set of practical interests mirrors a set, described previously, with which Greco himself fills out the Barney case.

bird in the tree. However, the singing that Tony hears is in fact that of a y-type bird, it just so happens that there is also an x-type bird in the same tree.

As in the Roddy case, in this Lost Birds case there is a true belief and there appears to be an exercise of cognitive ability. And, as to the Roddy case, to this Lost Birds case there is the response intuition that the protagonist doesn't know.

Lost Birds, case two:
On a separate occasion that day, Tony ventures up to the foot of a tree, listens carefully, hears singing, and forms the true belief that there is an x-type bird in the tree. The singing that Tony hears is in fact that of an x-type bird, it just so happens that in each of the five surrounding trees there is a y-type bird singing.

As in the Barney case, in the second Lost Birds case there is a true belief. And, as to the Barney case, to this case there is the response intuition that the protagonist doesn't know. My hope is that the cases taken together also show that intuitively it would be very odd to think that Tony has an environment-relative ability in one but not the other, or that Tony is exercising an environment-relative ability in one but not the other. I only need the latter to suggest a problem with Greco's handling of the cases.

It's possible, however, to press the point further in a way that reveals something interesting about Roddy-type cases at least and, if correct, implies that a defence of Greco's current way of handling the two cases differently is hopeless. Suppose there were just one x-type bird and one y-type bird in Tony's vicinity and they both happened to be in the same tree. In fact, Tony is listening at the foot of the tree. Let's say that the hundreds of other x-type birds and the four other y-type birds are off at the other end of Tony's part of the rainforest. Now if the y-type bird but not the x-type bird is singing, then it's a Roddy-type case. If it's the x-type bird but not the y-type bird that is singing, then it is a Barney-type case. The way in which Tony forms his belief is such that he could have easily been wrong. If the foregoing is accepted, then Greco's claim that there is an exercise of ability in Roddy-type cases but not in Barney-type cases is untenable. After all now a Roddy-type case, on Greco's analysis, should be seen as like the Barney-type case, except that in the Roddy case there

is no causal connection between the "unreliable cognitive faculty" (given the environment)/reliable ability to the truth of the belief in the Roddy case whereas there is in the Barney case.

6. *Greco: Possible responses*

How might Greco respond? Rather than alter how he handles the two cases, Greco might argue that at least one of the cases as described should itself be altered. Noting that Barney-type cases require for the target belief's basis to be such that the protagonist could have easily been wrong, Greco might argue as follows. Roddy-type cases have a necessary feature: although in fact the subject's belief basis is *not* appropriately connected to the belief's truth, it easily could have been. Whereas in the case as imagined, the subject's belief is lacking an appropriate connection to truth, it is not lacking such a connection in nearby possible worlds.

It's hard to see how this could be right in the case as described. Presumably in nearby worlds in which the sheep isn't, from Roddy's line of vision, standing directly behind the sheep shaped object but rather is standing a little to the side of the sheep shaped object, Roddy will form the false belief that there are *two* sheep in the field. Among the multitude of nearby cases, there is just one kind of variation in which Roddy's sheep belief is true due to Roddy's exercising a cognitive ability. In that variation, which is an instance of a Barney-type case, Roddy sees an actual sheep that stands not behind, but in front of, the sheep-shaped object that in the original case blocks Roddy's view of the actual sheep. Looking at an actual sheep, Roddy forms the true belief that there is *one* sheep in the field. But given the belief's basis, Roddy could easily have formed the false belief that there are two sheep in the field. When we think of it like this, the Roddy case, as originally described, seems to imply that his belief was formed in the case in such a way that he could have easily been wrong. Of course, in general, forming beliefs about the number of sheep in a field on the basis of perception is a good way of forming beliefs but then the same is true of forming beliefs about the presence of barns.[12]

12. Defenders of Greco's position might want to point to a more recent articulation of that position. Greco (2012, 18), however, doesn't revise his account of the structure of knowledge-relevant abilities. "S has a knowledge-relevant ability A (R/C/D) relative to an environment E = S has a disposition to believe truths in range R when in circumstances C and environment E, with degree of reliability D." Neither does Greco (2012, 22) alter his diagnosis of the Barney

It looks like my analysis of the Roddy case is applicable to other standard Gettier cases.[13] Consider the Edmund case as described by Pritchard (2012, 250) which is closely based on an original Gettier case:

> Edmund forms a belief that Jones owns a Ford on excellent grounds. He then validly infers that either Jones owns a Ford or Smith is in Barcelona, and accordingly forms a belief in this entailed proposition solely on the basis of his grounds for believing the entailing proposition and the relevant deduction. As it happens, the entailing proposition is false; the entailed proposition, however, is true since it just so happens (and unbeknownst to Jones) that Smith is in Barcelona.

Again, once the protagonist's belief being true is not explained by the protagonist's basis for belief in the particular case, then it will be vulnerable to the same charge. Although Edmund's basis for belief may generally be a good one, in this case it's not a good basis for believing the target proposition. It seems that Edmund's belief is only true because of the presence of intervening luck.

An interesting point to be taken from the cases is that if there is a requirement to assess whether an environment-relative ability is present in a particular case, then standard Gettier-type cases may be just as likely to receive a negative verdict as Barney-type cases. What's worse for Greco's position, given what he currently writes about the cases, such a way of looking at the cases makes standard Gettier-type cases appear to be epistemically worse than Barney-type cases. In general, however, this seems fitting given that Barney-type cases aren't quite so widely rejected as cases of knowledge, and that, perhaps relatedly, they seem closer to being cases of knowledge than standard Gettier cases.

The reason Greco generally has to resist saying that there is an exercise of ability in Barney-type cases, is to avoid the counterintuitive result that such cases are cases of knowledge. However, if one has a different analysis of knowledge, one that does not claim that knowledge is true belief because of the exercise of ability, then one can accept that there

---

case. Barney "believes from a disposition that is reliable relative to normal environments, but not relative to the environment he is in." In other words, Barney isn't believing from the relevant reliable ability.

13. Interestingly, it appears that what I've written does not apply to lottery-type cases. Whether one should consider lottery-type cases as cases of justified true belief is, however, debatable. See Smith (2010).

is true belief because of the exercise of ability in such cases without that committing one to saying that such cases are cases of knowledge. For example, if one were to claim that for an analysis of knowledge both an ability condition and an anti-luck condition are needed, rather than just an ability condition, then one could say that there is an exercise of ability in the Barney case without being committed to saying that it is a case of knowledge.

But isn't what Greco writes about abilities being environment-relative plausible? While it is plausible to deny that, say, Tiger Woods can exercise a reliable ability to sink putts in hurricane-force winds, the claim that ability implies reliability seems less intuitive. Plausibly, a long distance Olympic gold medalist's career best time is down to her ability; after all it's not as if non-athletes could luckily replicate her time, and yet it's precisely the kind of instance of an exercise of ability that is not reliable. It's not reliable in that even though we think the time is down to her ability we don't necessarily think that her ability is such that she can reliably reproduce it. While what a good alternative account of ability would look like requires spelling out, my point here is that, apart from the difficulties it leads Greco into outlined in this paper, there are independent, intuitive grounds to be wary of Greco's account.[14]

What's the dialectical upshot of my criticism of Greco's solution to the Gettier problem? If Greco were to accept my criticism of his solution, then it seems to me there is only one way for Greco to respond: to claim that, neither in Barney-type cases nor in standard Gettier cases, is there an exercise of the relevant cognitive ability. There may be some standard Gettier cases about which Greco can continue to hold that there is an exercise of ability, but so much the messier for his account.[15] Greco will no longer be able to provide the diagnosis of standard Gettier cases which was such an attractive feature of his account. Not only that, but it will also be unclear whether he'll even be in a position to say that cases that are currently taken to be standard Gettier cases are indeed Gettier cases. This is for the same reason that it appears that Greco might be committed to saying that Barney-type cases are not cases of justified true belief. While Greco need not abandon his account of knowledge as true belief because of the exercise of ability, I take the dialectical upshot of what I've argued for to be that his account is significantly less attractive than it initially appeared.

14. For alternative accounts of ability see Maier (2011).

15. Thanks to an anonymous reviewer for pointing out a Gettier case for which Greco could still provide his original diagnosis.

7. *Conclusion*

In this paper I argued that Greco's treatment of Barney-type cases is problematic in view of his other theoretical commitments. More specifically, I argued that he is committed to saying that in the Barney case there is not an exercise of the relevant cognitive ability, but that that claim appears unsustainable if he is to maintain his current analysis of standard Gettier-type cases. I did so by articulating and drawing on the Lost Birds cases. I argued that if we accept that there is no exercise of ability in the Barney-type Lost Bird case, then we're also committed to saying that there is no exercise of ability in the Roddy-type Lost Bird case.[16]

References

Chisholm, Roderick 1977: *Theory of Knowledge*, (2nd ed.). Englewood Cliffs, New Jersey: Prentice-Hall.

Craig, Edward 1990: *Knowledge and the State of Nature: An Essay in Conceptual Synthesis*. Oxford: Clarendon Press.

Goldman, Alvin I. 1976: "Discrimination and Perceptual Knowledge". *Journal of Philosophy* 73, 771–791.

Greco, John 2012: "A (Different) Virtue Epistemology". *Philosophy and Phenomenological Research* 85(1), 1–26.

— 2010: *Achieving Knowledge*. Cambridge: Cambridge University Press.

— 2009: "The Value Problem". In: Adrian Haddock, Alan Millar & Duncan Pritchard (eds.), *Epistemic Value*. Oxford: Oxford University Press, 313–321.

— 2009b: "Knowledge and Success from Ability". *Philosophical Studies* 142 (1), 17–26.

Greco, John & Turri, John, "Virtue Epistemology". *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/spr2011/entries/epistemology-virtue/>.

Maier, John, "Abilities". *The Stanford Encyclopedia of Philosophy (Fall 2011 Edition),* Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2011/entries/abilities/>.

16. I thank everyone who provided feedback on this paper. In particular I thank Duncan Pritchard, Andrea Kruse, Anthony Bolos, and several anonymous reviewers from this journal and another.

Pritchard, Duncan 2012: "Anti-Luck Virtue Epistemology". *The Journal of Philosophy* 109, 247–279.
— 2010: "Knowledge and Understanding". In Alan Haddock, Adrian Millar & Duncan Pritchard (eds.), *The Nature and Value of Knowledge: Three Investigations*. Oxford: Oxford University Press, 3–88.
Smith, Martin 2010: "What Else Justification Could Be". *Noûs* 44(1), 10–31.

Grazer Philosophische Studien
90 (2014), 213–230.

THE DISCRIMINATION ARGUMENT AND THE STANDARD STRATEGY

Simon DIERIG
University of Heidelberg

Summary
Two major objections have been raised to Boghossian's discrimination argument for the incompatibility of externalism and a priori self-knowledge. Proponents of the first objection claim that thoughts about "twin water" are not relevant alternatives to thoughts about water. Advocates of the second objection, the so-called standard strategy, argue with recourse to Burge's account of self-knowledge that the ability to rule out relevant alternatives is not required for knowledge. In this essay, it is shown that the standard strategy does not provide us with a convincing response to the discrimination argument.

I.

Boghossian was the first to advance an explicit argument for the incompatibility of semantic externalism and a priori self-knowledge, the so-called discrimination argument.[1] Following Warfield's reconstruction, it can be formulated as follows (see Warfield 1992, 234f.): To know a priori that p is the case, one has to be able to rule out a priori all relevant alternatives to p. But Oscar, our protagonist, cannot rule out a priori that he thinks that twater is wet. For if he were on twin earth, thinking that twater is wet, things would seem to him exactly as they seem to him in reality. (This claim is meant to follow from an externalist interpretation of the twin earth scenario.) Moreover, the proposition that Oscar thinks that twater is wet is a relevant alternative to the fact that he thinks that water

1. See Boghossian 1989, 12–14. The term "discrimination argument" is due to Brown (2004, 26). For further arguments for incompatibilism, see Bilgrami (1993, 240), Boghossian (1989, 22f.), Boghossian (1992, 18–22), Boghossian (1997, 165f.), Brown (1995, 152–155), Brown (2000, 118, 121 and 128), Brown (2004, 121 and 123), Brueckner (1990, 448), Brueckner (1994, 327f.), Butler (1997, 787f.) and McKinsey (1991, 15).

is wet. Therefore, Oscar does not know a priori that he thinks that water is wet.[2]

Two major objections have been leveled at this argument. According to the first objection, the proposition that Oscar thinks that twater is wet is not a *relevant* alternative to the fact that he thinks that water is wet.[3] The third premise of Boghossian's argument is therefore mistaken. Proponents of the second chief objection to the discrimination argument hold that the first premise of this argument is wrong. The ability to rule out relevant alternatives is, according to them, not necessary for knowledge. To support this thesis, they draw on Burge's account of self-knowledge (see Burge 1988, 1996). It is Burge's view that second-order thoughts about one's own thoughts are "self-verifying." From this it is concluded that thoughts about one's own thoughts are infallible and therefore amount to knowledge. Thus, our protagonist Oscar knows that he thinks that water is wet. But he cannot rule out all relevant alternatives to this thought. For he cannot rule out that he is on twin earth, thinking that twater is wet. Hence, the ability to rule out relevant alternatives is not necessary for knowledge.

This second objection to the discrimination argument has been called "the standard strategy" of criticizing Boghossian's argument because a number of philosophers think that it is an appropriate rejoinder to the argument in question.[4] In this essay, I attempt to show that these philosophers are mistaken. The standard strategy does not provide us with an adequate response to the discrimination argument. It founders for two reasons. First, one cannot invoke Burge's stance on self-knowledge in order to show that second-order thoughts are infallible or at least reliable. Second, neither infallibility nor reliability are sufficient for knowledge.

---

2. Here and in what follows I assume familiarity with Putnam's and Burge's twin earth thought experiments and the relevant-alternatives approach to knowledge developed by Dretske and Goldman. The classical sources for twin earth are Putnam (1975) and Burge (1979, 1982, 1986). For the relevant-alternatives account, see Dretske (1970) and Goldman (1976).

3. See Warfield 1992, 234f. For further discussion of this objection, see Ludlow (1995, 46–49), Brown (2004, 138–142), Goldberg (2006, 310f.), Gerken (2009, 124–131) and Dierig (2010, 75–78).

4. See Butler 1997, 780–783 and 790. Proponents of the standard strategy are, for example, Burge, Stalnaker, Falvey and Owens and Goldberg (see Burge 1988, Stalnaker 1990, Falvey and Owens 1994, Goldberg 2005, 2006).

II.

Burge claims that our thoughts about our own thoughts are self-verifying and therefore infallible: It is impossible that a certain person thinks that she thinks that p even though she does not think that p (see Burge 1988, 649). The second-order thought that one thinks that p is self-verifying because in thinking it one ipso facto thinks that p. In Burge's words: "When one knows that one is thinking that p, one is not taking one's thought (or thinking) that p merely as an object. One is thinking that p in the very event of thinking knowledgeably that one is thinking it. It is thought and thought about in the same mental act." (Burge 1988, 654)

Burge's theory of self-verification can be understood in (at least) two ways, depending on how his notion of a thought is interpreted. According to the first interpretation, thoughts in Burge's sense are what are commonly called "occurrent thoughts." If one occurrently thinks that p is the case, one inwardly says the sentence "p" to oneself. Taking this interpretation of Burge's notion of a thought as a basis, one can explain the infallibility of second-order thoughts as follows: If a certain person thinks that she thinks that p, she inwardly says the sentence "I think that p" to herself. But the sentence "p" is part of the sentence "I think that p." Thus, the person in question inwardly says the sentence "p" to herself. From this it can be inferred that she occurrently thinks that p. In short, if a certain person thinks that she thinks that p, she must think that p as well.

Second-order thoughts are also infallible according to a second interpretation of Burge's notion of a thought. Gibbons suggests to understand thinking in Burge's sense as the most general kind of propositional attitude: "For any propositional attitude, if you stand in that relation to a proposition, then you think that proposition." (Gibbons 2001, 19) Let us call this thesis "assumption A." According to Gibbons, a second thesis holds true for thinking in Burge's sense. Gibbons formulates it as follows: "If you think a proposition with propositional structure, you must think the propositional constituents." (Gibbons 2001, 19) Let us call this thesis "assumption B."

With the help of the two assumptions A and B, Gibbons constructs the following infallibility proof (see Gibbons 2001, 19f.): Suppose a certain person believes that she thinks that p. Falling back on the assumption A, it can be inferred that this person thinks that she thinks that p. From this claim and the assumption B it follows that the person in question thinks

that p. In brief, if one believes that one thinks that p, one has to think that p.

Our considerations so far show that we cannot err with regard to our own thoughts, provided the notion of a thought is understood in one of the two ways explained above. Burge's account of self-knowledge is therefore basically correct.[5] But what does this mean for Boghossian's discrimination argument? Apparently, Burge rejects this argument because of the ideas described above. And others, adherents of the "standard strategy," have followed him in doing so. However, neither Burge nor his followers have explained in more detail how an objection to the discrimination argument can be contrived on the basis of his account of self-knowledge.

Let me begin my attempt to reconstruct the standard strategy by considering how Bernecker understands Burge's account of self-knowledge (see Bernecker 1998, 338). In Bernecker's view, Burge's theory only shows that beliefs about one's own beliefs are necessarily true or infallible. But it does not show that these beliefs constitute *knowledge*.[6] According to Bernecker, one needs an additional premise in order to draw the conclusion that beliefs about one's own beliefs amount to knowledge, namely reliabilism with regard to knowledge. The classical version of reliabilism is to be found in Goldman's seminal article "Discrimination and Perceptual Knowledge" (see Goldman 1976, 771 and 785f.). His main claim in that essay is that a certain person knows that p iff the process which causes her belief that p is reliable, which is again the case iff there is no relevant counterfactual situation in which the process in question causes the belief that p and this belief is false.

Drawing on Goldman's reliabilism, the standard strategy can now be formulated as follows: Burge's theory of self-knowledge implies that our second-order thoughts are infallible and the processes which produce them therefore reliable in Goldman's sense. From this claim and the conten-

---

5. I have concentrated here on aspects of Burge's account which can be rationally reconstructed. Certain parts of his theory which are, to my mind, misguided—such as his claim that first-order thoughts are contained in second-order thoughts (see Burge 1988, 659f.) and that self-referential thoughts of the kind he considers are self-verifying (see Burge 1988, 649 and 658)—have been omitted from the start. For a criticism of Burge's first claim, see Sawyer (1999, 372f.); for a critique of his second contention, see Warfield (2005, 173–177).

6. See also Brueckner 1990, 451, fn. 11. Boghossian makes a very similar point with regard to Davidson's version of the standard strategy: It only shows that the beliefs in question are reliable, but not that they constitute knowledge (see Boghossian 1994, 35). Finally, Vahid maintains that although the standard strategy "makes the cognizer always *right* about what she thinks, it does not furnish us with enough grounds to attribute *knowledge* to her." (Vahid 2003, 376)

tion that reliability is sufficient for knowledge it can be inferred that our thoughts about our own thoughts amount to knowledge. Our protagonist, Oscar, therefore knows that he thinks that water is wet even though he cannot rule out the relevant alternative that he thinks that twater is wet. Hence, it can be concluded that the ability to rule out relevant alternatives is not necessary for knowledge.

One problem with this objection to the discrimination argument is the narrow range of Burge's account of self-knowledge. Boghossian has drawn attention to the fact that only thoughts about *one's own present thoughts* are in Burge's sense self-verifying (see Boghossian 1989, 20ff.). Thoughts with the content "I *believe* that water is wet" or the content "I *have just thought* that water is wet" are in Boghossian's opinion not self-verifying. The two infallibility proofs outlined at the beginning of this section cannot be carried out for them. But if Burge's theory of self-knowledge is restricted to thoughts about one's own present thoughts, it does not imply, for example, that one's beliefs about one's own beliefs are infallible or at least reliable. It appears, therefore, that the standard strategy is unfeasible for certain kinds of second-order thinking, like beliefs about one's own beliefs, and for this reason does not provide us with a full-fledged response to the discrimination argument.[7]

Yet a proponent of the standard strategy will not surrender that quickly. For he can correctly point out that Burge's account of self-knowledge amounts to more than his theory of self-verifying thoughts. I quote Burge: "If background conditions are different enough so that there is another object of reference in one's self-referential thinking, they are also different enough so that there is another thought." (Burge 1988, 659) Davidson summarizes this view of Burge's as follows: "what determines the contents of thoughts also determines what the thinker thinks the contents are." (Davidson 1988, 664) Unlike Burge's theory of self-verification, which obtains only for one's thoughts about one's own present thoughts, his account of content determination expressed in the two passages just quoted should apparently pertain for all kinds of first-order propositional attitudes.

Burge has not distinguished clearly enough between his views on self-verification and his account of the content determination of propositional

7. Contrary to what Warfield claims, Boghossian's observation is consequently not "irrelevant to the question of whether Burge has shown that privileged self-knowledge and externalism are compatible." (Warfield 2005, 173)

attitudes.[8] This may be the reason why some proponents of the standard strategy have claimed that his theory of content determination implies that our beliefs about our own beliefs are necessarily true (see, for example, Bernecker 1998, 338). Advocates of the standard strategy might go even further, alleging that it emerges from Burge's account of content determination that our thoughts about all our propositional attitudes are infallible or at least reliable. If this allegation is true, the standard strategy is not only viable for self-verifying thoughts, but for all kinds of second-order thinking.

But does it really follow from Burge's account of content determination that our beliefs about our own beliefs are infallible or at least reliable? To answer this question, it has to be settled first what Burge means with the expression "background conditions." I suggest the following interpretation: Background conditions in his sense are environmental conditions which are, according to externalists, part of the individuation conditions of propositional attitudes. Evidence for this interpretation is, amongst others, that he speaks of "*nonindividualistic* background conditions." (Burge 1988, 659, my emphasis)

Burge's theory of content determination expressed in the passage quoted above can now be formulated as follows (with regard to beliefs): If the environmental conditions in a counterfactual situation are different enough from the actual environmental conditions so that a certain person no longer has the belief that p (in this counterfactual situation), then the environmental conditions in this counterfactual situation are different enough from the actual environmental conditions so that the person in question no longer has the second-order belief that she believes that p (in this counterfactual situation).

Understood in this way, Burge's account of content determination can be established along the following lines: If at all, it can only be shown that a person no longer has the belief that p in a certain counterfactual situation because of a change in her environment by arguing that the person in question no longer has, due to her different surroundings, all the concepts contained in the content p. From this it can be concluded that if a person no longer has the belief that p in a certain counterfactual situation because the environmental conditions are different from those in the actual world, the person at issue no longer possesses, owing to her different surroundings, all the concepts contained in the content p. But

8. Sawyer emphasizes the importance of this distinction (see Sawyer 2002, 121f. and 126).

in order to have the second-order belief that one believes that p, one has to possess all the concepts incorporated in the proposition p. Thus, if a person no longer has the belief that p in a certain counterfactual situation because of changes in her environment, she no longer has the second-order belief that she believes that p. In brief, Burge's account of content determination is true for beliefs about one's own beliefs. Analogously, it can be argued that his account is also true for all other kinds of second-order thinking.[9]

Does it follow from what has just been shown for beliefs about one's own beliefs that they are infallible or at least reliable? Let me ask a related question first: Does it follow from Burge's thesis concerning content determination that our protagonist Oscar does not wrongly believe on twin earth that he believes that water is wet? That this question has to be answered in the affirmative can be seen as follows: Either individualism is true or externalism is true. If individualism is true, then Oscar believes on twin earth, as he does on earth, that water is wet. Therefore, if individualism is true, Oscar does not mistakenly believe on twin earth that he believes that water is wet. If externalism is true, then our protagonist does not have the belief that water is wet on twin earth because the environmental conditions on twin earth are distinct from those on earth. Given Burge's theory of content determination, it follows that the environmental conditions on twin earth are different enough from those on earth so that Oscar no longer believes on twin earth that he believes that water is wet. Again it can be concluded that he does not wrongly believe on twin earth that he believes that water is wet.

To establish that it emerges from Burge's account of content determination that Oscar's second-order belief is infallible it is, however, not enough to show that Burge's account implies that Oscar does not wrongly believe *on twin earth* that he believes that water is wet. Rather it has to be demonstrated that Burge's theory implies that there is *no* counterfactual situation in which Oscar mistakenly believes that he believes that water is wet.

9. Here is a second way to substantiate Burge's account of content determination: He holds that first-order thoughts are externally determined and that their contents are contained in the contents of the corresponding second-order thoughts (see Burge 1988, 659f.). From this one can gather that the same external conditions which determine first-order thoughts also contribute to the determination of the corresponding second-order thoughts. Thus, Burge's doctrine of content determination is true. (For the first step of this argument, see Bernecker (1998, 338); for the second step, see Wright (1992, 76) and Sawyer (1998, 524).)—In contrast to the line of reasoning just put forward, my argument for Burge's doctrine has the merit of not relying on an externalist understanding of propositional attitudes.

Consider a counterfactual situation in which Oscar does not believe that water is wet, as is the case in actuality, but rather he believes that London is pretty because he no longer has the disposition to sincerely assent to the sentence "Water is wet," but rather he has the disposition to sincerely assent to the sentence "London is pretty." In this counterfactual situation, Oscar does not believe that water is wet, but the reason for this is not that the environmental conditions are different from those in the actual world, but rather that he does not have a certain disposition which he possesses in reality. One cannot therefore employ Burge's doctrine of content determination in order to show that Oscar does not have the second-order belief in the counterfactual situation in question. Thus, Burge's theory is compatible with the assumption that our protagonist has the wrong belief that he believes that water is wet in the counterfactual situation described above. In other words, Burge's account of content determination does not imply that Oscar's second-order belief is infallible.

But does Burge's account perhaps imply the weaker claim that the process which causes Oscar's second-order belief is reliable? Let us begin to answer this question by looking again at the counterfactual situation outlined in the last paragraph. We have seen that there is nothing in Burge's account that rules out that Oscar has the wrong second-order belief that he believes that water is wet in this counterfactual situation. Since Burge's theory addresses neither the relevancy of counterfactual situations nor the causal ancestry of second-order beliefs, one can strengthen this claim as follows: Burge's account does not rule out that the process which causes Oscar's second-order belief in reality also causes this belief in the, relevant, counterfactual situation in question, in which this belief is wrong. Burge's account is therefore compatible with the assumption that the process which produces Oscar's second-order belief in reality also produces this belief in a relevant counterfactual situation in which this belief is false. But this means that it does not follow from Burge's account that the process which brings about Oscar's second-order belief in the actual world is reliable. Generalizing from the example of Oscar, one arrives at the conclusion that Burge's theory of content determination implies neither that our beliefs about our own beliefs are infallible nor that the processes which produce them are reliable.

To sum up, according to Burge's account of self-verification, thoughts about one's own present thoughts are self-verifying and therefore infallible. But the main part of our second-order attitudes does not consist in thoughts about our own present thoughts. Burge's theory of self-verification

does not imply that second-order states of this kind—and, in particular, beliefs about one's own beliefs—are self-verifying and, in consequence, infallible. The standard strategy thus threatens to fail if there is no other way to show that second-order attitudes which are not thoughts about one's own present thoughts are infallible or at least reliable. At first glance, Burge's account of content determination seems to provide a solution. In contrast to his theory of self-verification, it is true for all kinds of second-order attitudes. But it has been argued that it does not follow from Burge's account of content determination that beliefs about one's own beliefs are infallible or at least reliable. The standard strategy is therefore not viable for all kinds of second-order thinking—in particular, not for beliefs about one's own beliefs—and is accordingly not acceptable as a general response to the discrimination argument.

III.

The standard strategy in the form reconstructed in the last section presupposes not only the claim that Burge's account of self-knowledge implies that our second-order thoughts are infallible or at least reliable, but also the contention that reliability is sufficient for knowledge. In this section, it is argued that the latter contention, like the former claim, does not stand up to closer scrutiny. In a second step, I will examine some suggestions for revising the standard strategy in such a way that it no longer presupposes a reliabilist account of knowledge.

Let us start by looking at an example which Gibbons offers to show that infallibility is, even when it comes to second-order thinking, not sufficient for knowledge (see Gibbons 2001, 22f.). Gibbons writes: "Suppose an insecure student, Harry, has read Descartes and knows that thinking is the most general propositional attitude. (…) Unfortunately, Harry also believes that in order to understand a proposition, you have to grasp it through the natural light of reason. Harry tries to grasp a proposition through the natural light of reason. But whatever else goes on, he is never aware of any light, natural or otherwise. Harry begins to suspect that he is incapable of thinking. (…) Ashamed at his imagined disability, Harry begins daydreaming about how nice it would be to think. Sometimes during these reveries Harry believes for a moment that he really is thinking. But he soon dismisses these second-order beliefs as wishful thinking (well, not exactly wishful *thinking* …)." (Gibbons 2001, 22)

Intuitively, the second-order beliefs just mentioned do not amount to knowledge. This intuition can be substantiated by the following reasons: First, the beliefs in question occur during a daydream. Second, they are soon dismissed as wishful thinking. Third, Harry holds views which imply that he does *not* think the first-order propositions in question. These views are: In order to think a proposition, one has to grasp it through the natural light of reason. But in trying to grasp propositions, I am never aware of any light.

The three lines of argument just outlined indicate that Harry's second-order beliefs do not constitute knowledge. But Harry's beliefs are beliefs about his own present thoughts. With the aid of Burge's theory of self-knowledge, it can therefore be concluded that they are infallible. Thus, even when it comes to second-order thinking, infallibility is not sufficient for knowledge. Since infallibility implies reliability, it can be inferred, moreover, that reliability is not sufficient for knowledge either. But this means, once more, that the standard strategy in the form reconstructed above is untenable.

Given this result, the question arises whether the standard strategy can perhaps be rephrased in such a way that it does not presuppose such a controversial doctrine as reliabilism with regard to knowledge. According to Goldberg, thoughts about one's own present thoughts are not only infallible in virtue of the mechanism of self-verification and therefore objectively justified to the highest possible degree, but one can recognize this by reflection (see Goldberg 2005, 141f.). Goldberg claims that if our second-order thoughts enjoy the highest possible degree of objective justification *and we can recognize this by reflection*, we have knowledge of our first-order thoughts (see Goldberg 2005, 142–144.). If one substitutes this claim for the contention that reliability is sufficient for knowledge, one gets a version of the standard strategy which is not purely reliabilist and which for this reason may be thought to be more promising than the original version of the standard strategy.

It turns out on closer examination, however, that a reconstruction of the standard strategy in Goldberg's style is not doing any better than a reconstruction which falls back upon reliabilism. For as the example of Harry, the insecure student, makes plain, even *internally accessible* infallibility (or objective justification to the highest possible degree) is not sufficient for knowledge. Harry's second-order beliefs are infallible. And he can recognize this by reflection alone. In other words, the fact that his second-order beliefs are infallible is internally accessible. But Harry's beliefs

do not amount to knowledge (for the reasons given above). Hence, even *internally accessible* infallibility is not a sufficient condition for knowledge. Goldberg's version of the standard strategy fails.

Let us consider a second attempt to formulate the standard strategy without drawing on a reliabilist account of knowledge. According to some internalists in epistemology, only internally accessible objective justification *together with subjective justification* is sufficient for knowledge. A belief of a person is meant to be subjectively justified iff the person in question bases her belief on a certain justification for it (see Goldberg 2005, 141, fn. 8, and 143). Can the standard strategy be rescued by substituting the internalist claim just mentioned for the assumption that reliability is sufficient for knowledge?

Note, first of all, that the example of Harry cannot be employed to show that internally accessible objective plus subjective justification is not sufficient for knowledge. Wright has drawn attention to the fact that it is, cases of "self-interpretation" aside, always inappropriate to ask for a reason for a self-ascription of a propositional attitude (see Wright 1998, 14ff.). From this it emerges that we do not have reasons for our second-order beliefs about our own thoughts. But if Harry does not have a reason for his second-order beliefs, he cannot base these beliefs on a reason or justification for them. In other words, his second-order beliefs are not subjectively justified. It follows that the case of Harry is not a counterexample to the claim that internally accessible objective plus subjective justification is sufficient for knowledge.

The observation that one does not have reasons for one's beliefs about one's own propositional attitudes can be used to defend the claim just mentioned against alleged counterexamples, as the case of the insecure student Harry. But it can also be used to attack the non-reliabilist version of the standard strategy currently under discussion. One of the premises of this version of the standard strategy, when fully stated, is the claim that Oscar has subjective justification for his belief that he thinks that water is wet. But if one does not have reasons for one's second-order beliefs, this premise is wrong. Thus, even when the claim that internally accessible objective plus subjective justification is sufficient for knowledge should turn out to be true, the non-reliabilist variant of the standard strategy we are considering now is unsound.[10]

10. In one passage in "Individualism and Self-Knowledge" Burge seems to claim that one knows one's own thoughts simply by self-ascribing them correctly (see Burge 1988, 656). If one substitutes this claim for the contention that reliability is sufficient for knowledge, one gets

It has been argued that the original version of the standard strategy founders because Burge's account of self-knowledge does not imply that our second-order thoughts are infallible or at least reliable, and also because reliability is, even when it comes to second-order thinking, not sufficient for knowledge. In view of these shortcomings of the standard strategy, I have discussed endeavors to reformulate it without falling back upon a reliabilist account of knowledge. But both non-reliabilist variants of the standard strategy which have been examined have proven to be untenable.

IV.

In my attempts to reconstruct the standard strategy I have so far employed Goldman's notion of reliability as he analyzes it in his article "Discrimination and Perceptual Knowledge." Goldman claims in that essay that a process which causes a belief that p is reliable iff there is no relevant counterfactual situation in which the process in question causes the belief that p and this belief is false (see Goldman 1976, 771 and 785f.). In a later article, Goldman introduces a notion of reliability which differs considerably from his earlier conception. According to his new analysis, a process is reliable iff the probability is high that the process in question produces true beliefs (see Goldman 1979, 10f.).

This notion of reliability, known as the concept of "global reliability," differs in two important respects from the notion of reliability expounded in "Discrimination and Perceptual Knowledge," a notion which is known as the concept of "local reliability."[11] First, whereas the answer to the question whether a process is locally reliable depends on whether this process causes a wrong belief in certain *counterfactual situations*, the answer to the question whether a process is globally reliable depends on its statistical properties *in the actual world*. Second, a process which causes the belief that p is locally reliable if there is no relevant counterfactual situation in which this process brings about the false belief *that p*. *Only the content p*

---

another non-reliabilist version of the standard strategy. Unfortunately for the compatibilist, Burge's claim just mentioned can be rebutted by the example of Harry, the insecure student. For further objections to Burge's claim, see Goldberg (2000) and McCullagh (2002).

11. See McGinn 1984, 536, and Goldman 1986, 44f. The distinction between "local" and "global reliability" is due to McGinn and Goldman; my usage of these expressions follows Brown's explanation of them (see Brown 2004, 120f.).

is pertinent to the local reliability of such a process. By contrast, not only the content p, but also *contents different from p* are pertinent to the global reliability of a process which causes the belief that p: A process of this kind is globally reliable if the probability is high that this process produces true beliefs; and the contents of these beliefs can of course be different from the content p.

Given that the concepts of local and global reliability differ in important ways, proponents of the standard strategy might think it promising to reformulate their objection to the discrimination argument by using the notion of global reliability, rather than the notion of local reliability. Is such a reformulation of the standard strategy successful? This question has to be answered in the negative for it can be shown that global reliability is, like local reliability, not sufficient for knowledge.

Consider, once more, the example of Harry, the insecure student. Is the process which causes his second-order beliefs globally reliable? Is the relative frequency, that is, probability, high that this process produces true beliefs? To answer this question, it has to be settled first how the process which causes Harry's second-order beliefs must be described. If it is accurately described as "introspection," it is globally reliable. For the process called "introspection" in most cases leads to true beliefs.

It might be objected that the process which brings about Harry's second-order beliefs should rather be described more specifically as "introspection of someone who believes that thinking a proposition implies that one grasps it through the natural light of reason." My reply to this objection is that even if the process which causes Harry's second-order beliefs is correctly described in the way suggested, it is globally reliable. This claim can be substantiated as follows: Unlike normal introspection which produces second-order beliefs in most cases of first-order propositional attitudes, introspection of the kind just characterized is, as it were, truncated because it often does not give rise to a second-order belief even though the corresponding first-order attitude is present, the reason for this being that no natural light of reason is perceived. However, introspection of the truncated kind does not cause second-order beliefs when the corresponding first-order attitude is missing. In other words, it does not bring about *mistaken* second-order beliefs about one's own first-order attitudes, at least not more often than normal introspection. Thus, not only regular introspection gives rise to true beliefs in most cases, but also the introspection of someone who believes that thinking a proposition implies that one grasps

it through the natural light of reason.[12] In brief, even if the process which causes Harry's second-order beliefs has to be described as the "irregular" introspection just mentioned, it is globally reliable.

Since Harry's second-order beliefs do not amount to knowledge (for the reasons given in section three), it can be inferred that global reliability is, even when it comes to second-order thinking, not sufficient for knowledge.[13] But this means that the standard strategy cannot be rescued by reformulating it in terms of global, rather than local, reliability.[14]

---

12. At this juncture, one might argue that introspection as just characterized admittedly does not cause false beliefs to the effect that one has a certain propositional attitude, but that it brings about wrong beliefs to the effect that one does *not* have a certain thought. To counter this objection, it needs to be emphasized that the process by which Harry arrives at his wrong beliefs that he does not have certain thoughts is different from the process—introspection of the truncated kind—by which he arrives, during his daydreams, at his second-order beliefs that he has certain thoughts. For whereas the former process involves *reasoning* concerning the natural light of reason, its connection to thinking and the fact that it is missing in particular cases, no such reasoning—indeed, no reasoning at all—is part of the process which causes Harry's second-order beliefs during his reveries.

13. One may raise the objection that the claim that global reliability is sufficient for knowledge can be justified by drawing on Goldman's famous barn example (see Goldman 1976, 772–775 and 785–787). After all, his argument for reliabilism, which is based on this example, is known as the standard argument for a reliabilist account of knowledge. Brown, for instance, calls it the "classic argument for reliabilism" (Brown 2004, 60). In response to this challenge, I want to point out that what the barn example shows is at most that global reliability is *necessary* for knowledge, but not that it is *sufficient* for it. Goldman distinguishes two versions of the barn scenario. What needs to be explained, according to him, is that his protagonist does not have a certain piece of knowledge in its second version. (In the following, I am only concerned with the second variant of the barn example.) The assumption that reliability (either local or global) is necessary for knowledge can be used to explain that the belief in question does not amount to knowledge because this assumption implies, together with the true claim that the relevant process is not reliable, that the belief in question does not constitute knowledge. In contrast, the contention that reliability is sufficient for knowledge cannot be used to explain that the belief in question does not amount to knowledge because this contention does not imply, together with the true claim that the relevant process is not reliable, that the belief in question does not constitute knowledge. Hence, the barn example shows at most that reliability (either local or global) is *necessary* for knowledge, but not that it is *sufficient* for it.

14. A third well-known notion of reliability, in addition to the concepts of local and global reliability, is due to Nozick (see Nozick 1981, 172–178). A reformulation of the standard strategy in terms of his conception of reliability is, however, not very promising since it has been argued persuasively that reliability in Nozick's sense is not sufficient for knowledge (see McGinn 1984, 532f.).

V.

The standard strategy, as it is usually understood and as I have understood it so far in this essay, is an objection to the claim that the ability to rule out relevant alternatives is necessary for knowledge. Strictly speaking, the first premise of the discrimination argument is, however, not tantamount to this claim, but rather amounts to the contention that the ability to rule out *a priori* relevant alternatives is necessary for *a priori* knowledge.

An opponent of the discrimination argument may therefore suggest reformulating the standard strategy as follows: It emerges from Burge's theory of self-knowledge that the processes which cause one's second-order thoughts are reliable. From this claim and the contention that reliability is, when it comes to second-order thinking, sufficient for a priori knowledge, it can be concluded that one's thoughts about one's own thoughts amount to a priori knowledge. Our protagonist Oscar therefore knows a priori that he thinks that water is wet even though he cannot rule out a priori the relevant alternative that he thinks that twater is wet. Thus, the ability to rule out a priori relevant alternatives is not necessary for a priori knowledge.

Both pivotal premises of this objection to the discrimination argument are mistaken. First, it has been shown that Burge's account of self-knowledge does not imply that the processes which produce our second-order thoughts are reliable. Second, if reliability is, when it comes to second-order thinking, not sufficient for knowledge, as has been argued in this essay, then reliability is, in the same area, a fortiori not sufficient for *a priori* knowledge.

References

Bernecker, Sven 1998: "Self-Knowledge and Closure". In: Peter Ludlow and Norah Martin (eds.), *Externalism and Self-Knowledge*. Stanford: CSLI Publications (= CSLI Lecture Notes 85), 333–349.

Bilgrami, Akeel 1993: "Can Externalism Be Reconciled with Self-Knowledge?". *Philosophical Topics* 20/1, 233–267.

Boghossian, Paul A. 1989: "Content and Self-Knowledge". *Philosophical Topics* 17/1, 5–26.

Boghossian, Paul A. 1992: "Externalism and Inference". In: Enrique Villanueva (ed.), *Rationality in Epistemology*. Atascadero: Ridgeview Publishing Company (= Philosophical Issues 2), 11–28.
— 1994: "The Transparency of Mental Content". In: James E. Tomberlin (ed.), *Logic and Language*. Atascadero: Ridgeview Publishing Company (= Philosophical Perspectives 8), 33–50.
— 1997: "What the Externalist Can Know A Priori". *Proceedings of the Aristotelian Society* 97, 161–175.
Brown, Jessica 1995: "The Incompatibility of Anti-Individualism and Privileged Access". *Analysis* 55, 149–156.
— 2000: "Reliabilism, Knowledge, and Mental Content". *Proceedings of the Aristotelian Society* 100, 115–135.
— 2004: *Anti-Individualism and Knowledge*. Cambridge, Mass., and London: The MIT Press (= Contemporary Philosophical Monographs 4).
Brueckner, Anthony 1990: "Scepticism about Knowledge of Content". *Mind* 99, 447–451.
— 1994: "Knowledge of Content and Knowledge of the World". *The Philosophical Review* 103, 327–343.
Burge, Tyler 1979: "Individualism and the Mental". In: Peter A. French, Theodore E. Uehling, Jr., and Howard K. Wettstein (eds.), *Studies in Metaphysics*. Minneapolis: University of Minnesota Press (= Midwest Studies in Philosophy 4), 73–121.
— 1982: "Other Bodies". In: Andrew Woodfield (ed.), *Thought and Object. Essays on Intentionality*. Oxford and New York: Oxford University Press, 97–120.
— 1986: "Intellectual Norms and Foundations of Mind". *The Journal of Philosophy* 83, 697–720.
— 1988: "Individualism and Self-Knowledge". *The Journal of Philosophy* 85, 649–663.
— 1996: "Our Entitlement to Self-Knowledge". *Proceedings of the Aristotelian Society* 96, 91–116.
Butler, Keith 1997: "Externalism, Internalism, and Knowledge of Content". *Philosophy and Phenomenological Research* 57, 773–800.
Davidson, Donald 1988: "Reply to Burge". *The Journal of Philosophy* 85, 664–665.
Dierig, Simon 2010: "The Discrimination Argument Revisited". *Erkenntnis* 72, 73–92.
Dretske, Fred I. 1970: "Epistemic Operators". *The Journal of Philosophy* 67, 1007–1023.
Falvey, Kevin, and Owens, Joseph 1994: "Externalism, Self-Knowledge, and Skepticism". *The Philosophical Review* 103, 107–137.
Gerken, Mikkel 2009: "Conceptual Equivocation and Epistemic Relevance". *Dialectica* 63, 117–132.

Gibbons, John 2001: "Externalism and Knowledge of the Attitudes". *The Philosophical Quarterly* 51, 13–28.
Goldberg, Sanford 2000: "Externalism and Authoritative Knowledge of Content: A New Incompatibilist Strategy". *Philosophical Studies* 100, 51–79.
— 2005: "The Dialectical Context of Boghossian's Memory Argument". *Canadian Journal of Philosophy* 35, 135–148.
— 2006: "Brown on Self-Knowledge and Discriminability". *Pacific Philosophical Quarterly* 87, 301–314.
Goldman, Alvin I. 1976: "Discrimination and Perceptual Knowledge". *The Journal of Philosophy* 73, 771–791.
— 1979: "What Is Justified Belief?". In: George S. Pappas (ed.), *Justification and Knowledge. New Studies in Epistemology*. Dordrecht, Boston and London: D. Reidel Publishing Company (= Philosophical Studies Series in Philosophy 17), 1–23.
— 1986: *Epistemology and Cognition*. Cambridge, Mass., and London: Harvard University Press.
Ludlow, Peter 1995: "Externalism, Self-Knowledge, and the Prevalence of Slow Switching". *Analysis* 55, 45–49.
McCullagh, Mark 2002: "Self-Knowledge Failures and First Person Authority". *Philosophy and Phenomenological Research* 64, 365–380.
McGinn, Colin 1984: "The Concept of Knowledge". In: Peter A. French, Theodore E. Uehling, Jr., and Howard K. Wettstein (eds.), *Causation and Causal Theories*. Minneapolis: University of Minnesota Press (= Midwest Studies in Philosophy 9), 529–554.
McKinsey, Michael 1991: "Anti-Individualism and Privileged Access". *Analysis* 51, 9–16.
Nozick, Robert 1981: *Philosophical Explanations*. Oxford: Clarendon Press.
Putnam, Hilary 1975: "The Meaning of 'Meaning' ". In: Hilary Putnam, *Mind, Language and Reality. Philosophical Papers, Volume 2*. Cambridge, New York and Melbourne: Cambridge University Press, 215–271.
Sawyer, Sarah 1998: "Privileged Access to the World". *Australasian Journal of Philosophy* 76, 523–533.
— 1999: "An Externalist Account of Introspective Knowledge". *Pacific Philosophical Quarterly* 80, 358–378.
— 2002: "In Defence of Burge's Thesis". *Philosophical Studies* 107, 109–128.
Stalnaker, Robert 1990: "Narrow Content". In: C. Anthony Anderson and Joseph Owens (eds.), *Propositional Attitudes. The Role of Content in Logic, Language, and Mind*. Stanford: Stanford University Press (= CSLI Lecture Notes 20), 131–145.
Vahid, Hamid 2003: "Externalism, Slow Switching and Privileged Self-Knowledge". *Philosophy and Phenomenological Research* 66, 370–388.

Warfield, Ted A. 1992: "Privileged Self-Knowledge and Externalism Are Compatible". *Analysis* 52, 232–237.
— 2005: "Tyler Burge's Self-Knowledge". *Grazer Philosophische Studien* 70, 169–178.
Wright, Crispin 1992: "On Putnam's Proof that We Are Not Brains-in-a-Vat". *Proceedings of the Aristotelian Society* 92, 67–94.
— 1998: "Self-Knowledge: The Wittgensteinian Legacy". In: Crispin Wright, Barry C. Smith and Cynthia Macdonald (eds.), *Knowing Our Own Minds*. Oxford and New York: Oxford University Press, 13–45.

Grazer Philosophische Studien
90 (2014), 231–251.

INTELLECTUALISM AGAINST EMPIRICISM*

Federico CASTELLANO
National University of Córdoba
National Scientific and Technical Research Council

Summary

Intellectualism is the philosophical view that thinking involves the activity of reason-giving. In this paper I argue that the intellectualist point of view is incompatible with any form of empiricism. First, I show that Traditional Empiricism collapses because it brings together two conflicting theses: the intellectualist thesis, according to which the normative properties of thoughts depend (rest) upon the activity of reason-giving, and the intuitive empiricist thesis, according to which the normative properties of empirical thoughts derive from perceptual experience. Second, I argue that McDowell's Minimal Empiricism collapses as well because of his attempt to make sense of an over-intellectualized and contradictory variety of empiricism: one that preserves both an intellectualist approach to thought and a conceptual but passive approach to perceptual experience.

Introduction

There is a well-known philosophical tradition called Intellectualism that claims thinking essentially involves the activity of reason-giving (Brewer 1999, 2005, Brandom 1994, 2002b, 2010, Davidson 1982, 1997, McDowell 1994, 2009a, 2009b, Sellars 1991).[1] According to this view, thinking

* Two previous versions of this paper were presented at the II Workshop on Concepts and Perception (Córdoba 2012) and the SADAF Colloquium (Rosario 2013). I want to thank all participants in the discussions, especially Juan Durán, Sean Kelly, and Pierre Steiner for very helpful comments. I also want to thank the "Grupo de Conceptos" (especially, Mariela Aguilera, Laura Danón, and Daniel Kalpokas) and the anonymous referee for encouraging comments and useful suggestions on a previous draft.

1. Although the concept "Intellectualism" has a long history in philosophy, I am concerned here with a particular interpretation according to which Intellectualism is the philosophical tradition embracing that thinking necessarily involves the ability to make moves within what Sellars has called "the logical space of reasons" (Sellars 1991, 169).

is a cognitive ability that one exercises reflectively by taking thoughts into account in reasoning. In this paper I will argue that the intellectualist point of view, frequently associated with Kant's critical philosophy, is incompatible with any form of empiricism. In particular, I shall argue that both Traditional[2] and McDowell's Minimal Empiricism (1994, 2009a) collapse because they support two conflicting theses: (a) the intellectualist thesis, according to which the normative properties of thoughts *depend* (*rest*) upon the activity of reason-giving, and (b) the intuitive empiricist thesis, according to which the normative properties of empirical thoughts *derive* from perceptual experience. Two ideas will become clear from my argumentation: first, Sellars's myth of the Given, which represents the most profound and powerful critique against Traditional Empiricism, is just a symptom of a widespread disease caused by bringing together these two theses. Second, McDowell's Minimal Empiricism collapses because of his attempt to make sense of an over-intellectualized and contradictory picture of empiricism, i.e., to make the intellectualist approach to thought compatible with a conceptual but passive and, consequently, non-intellectualist approach to perceptual experience.[3] It is important to make explicit that I will not argue directly against Intellectualism. What I am primarily concerned with here is the predicament those with an empiricist spirit find themselves in when following the intellectualist approach.

1. *The intellectualist approach to thought*

Intellectualism is the philosophical view that thinking essentially involves the ability to play a role in what Brandom—paraphrasing Sellars—has called "the game of giving and asking for reasons" (Brandom 2002b, 349).[4]

2. By "Traditional Empiricism" I mean what McDowell explicitly suggests, i.e., the theory that "answers the question 'Does empirical knowledge have a foundation?' … with an unqualified 'Yes'" (McDowell 2009e, 221). Under this notion, both Classical (British) and Logical Empiricism are included.

3. By "non-intellectualist" I mean that the normative properties of perceptual contents do not rest, unlike those of thoughts, upon reasons.

4. It is widely accepted that Intellectualism has its roots in Kant's and Hegel's idealism (see Brandom 1994, 2002a, McDowell 2009c, 2009d). Currently, however, Intellectualism is thought to be a comprehensive philosophical approach that includes a variety of ideas and theories regarding knowledge, language, and cognition, which are not always consistent with each other. In the next section I will present two conflicting lines within the core of the intellectualist tradition: a full-blooded holistic line and an empiricist line.

According to this position, thinking thoughts (or making judgments) is a cognitive ability that one *ultimately* exercises in a reflective way by taking them into account in reasoning.[5] The centrality of this approach rests on the widely accepted idea that in order to think a thought (or make a judgment), the ability to bring into play the reasons supporting the thought (or judgment) in question is required. To be precise, to make intelligible the idea of a subject as entertaining a thought, it is required—in the intellectualist view—that she be able to take into account the reasons in virtue of which she responds as she does. If she could not bear in mind such reasons—intellectualists argue—, then she is not thinking at all. Accordingly, a creature cannot think a thought unless she has the ability to grasp and bring the reasons (i.e., the reasons in virtue of which she responds as she does) into play in reasoning.

Upon consideration of this suggestion, an important question arises: what is "to grasp and bring the reasons into play in reasoning" supposed to mean?[6] Russell's Principle—revitalized and defended mostly by Evans (1982)—will provide some clues to the answer. This principle specifies that in order to think a thought, a certain kind of knowledge (or understanding) of its components is required.[7] According to this, in order to

5. I highlight "ultimately" because intellectualists certainly do not require that in order to have a thought, one necessarily has to reflect about the reasons that support it. It is enough that, if asked, one could (cf. Brandom 2002b, 352, and McDowell 2009b, 129).

6. One could suppose that the intellectualist approach involves some sort of reflexive explanation, since the ability to think requires the ability to grasp reasons, which is no more than the ability to think about thoughts. In general, intellectualists adhere to this idea. In their view, the ability to think a thought involves reflexivity in that the ability to think "first-order" thoughts implies the ability to think "second-order" thoughts—i.e., thoughts about thoughts (Bermúdez 2010, Davidson 1982, 1983, 1989, 1997). Davidson is probably one of the major representatives of the reflexive approach to thought. According to Davidson, a necessary condition for thinking is having the concept of "thought" because—following Davidson—the idea of having, for instance, the thought "something is a cat" makes no sense unless "you can make sense of the idea … of believing or judging that something is a cat which is not a cat" (Davidson 1997, 124). Although most intellectualists agree with such reflexive or meta-representational approach to thought, there is an intellectualist, though not totally reflexive, approach that is capturing great interest in current philosophy. I am referring to Brandom's *pragmatic* approach. In effect, Brandom defends that thinking involves the ability to grasp and bring reasons into play in reasoning. Nevertheless, he argues that grasping reasons is something we firstly do, not explicitly, but implicitly through social practices (see Brandom 1994, chap. 1; 2010).

7. Russell's principle states: "[*e*]*very proposition which we can understand must be composed wholly of constituents with which we are acquainted*" (Russell 1912, 58). It is well-known that this principle plays a fundamental role in Evans's *The Varieties of Reference* (1982). Nevertheless, Evans is not concerned with the semantic but with a cognitive interpretation of this principle. Although the principle was originally intended to specify which kind of knowledge is required

think the thought that *a* is F, it is required that the subject *masters* both the concepts *a* and F (cf. Evans 1982, chap. 4). As a result, if a subject is credited with the thought that snow is white, then—according to Russell's Principle—she must be also credited with mastering the relevant concepts "snow" and "white". It is important to note, however, that lacking one of the relevant concepts involved in a thought is enough to not be able to think (entertain) it. Briefly speaking, if a subject failed to master either the concept "snow" or the concept "white", then she would be unable to think the thought that snow is white. Consequently, a creature cannot think conceptual thoughts if she does not master *all* the relevant concepts involved in such thoughts.[8]

At this point, however, another question is inevitable: what does "to master the relevant concepts" mean? And here intellectualists like to say: to master a concept F is just to grasp (understand) the norm that prescribes whether or not something falls under F, and to grasp (understand) a norm is just to *follow* a rule correctly. Let me clarify this point.

It might be suggested that following rules is no more than acting in accordance with them. According to this interpretation, mastering a concept F must be read as having the ability to apply F according to a rule—a rule that prescribes whether something falls under F or not. Although such a picture of conceptual abilities may be intuitive, intellectualists point out that having the ability to act in accordance with a rule is not sufficient to master a concept—yet it is necessary. One could assume that non-human animals, primitive organisms, and even simple mechanisms

---

to *make a judgment*, Evans's interpretation is primarily concerned with what kind of cognitive ability is required for *thinking singular conceptual thoughts* (cf. Evans 1982, chap. 4). I am not totally convinced of Evans's interpretation. In my opinion, Russell's principle is not only intended to specify which kind of knowledge is required to think singular thoughts (i.e., thoughts containing singular terms or "referring expressions") but also to specify the cognitive ability required to think all kind of thoughts—whether they be singular or general (i.e., thoughts containing "general concepts"). In what follows, I will interpret Russell's principle in this broad sense.

8. I am neutral regarding the question "what ability do we learn first: thinking thoughts or mastering concepts?" What I am concerned with here is the conceptual dependence between the ability to think thoughts and the ability to master the concepts involved in such thoughts. In a relevant sense, having thoughts involves mastering the concepts involved in such thoughts; but in another sense, mastering concepts involves being able to think thoughts—since concepts are such if and only if they are exercised in thoughts. I do not want to discuss whether concepts or thoughts are first. Nevertheless, it is important to highlight that, since the ability to think necessarily involves the ability to master concepts and vice versa, an elucidation of the cognitive resources involved in mastering concepts must throw light on the cognitive resources involved in thinking.

behave intelligently since their behaviors are carried out in accordance with rules and, therefore, it seems as if they follow rules accurately. A mouse fleeing, for instance, is acting in accordance with a specific rule: the rule that orders it to flee whenever it is presented with danger. A sunflower moving its flowers toward the sun is acting in accordance with a rule as well: the rule that orders it to move its flowers toward the sun whenever it is presented with rays of sunlight. If we commit ourselves to the interpretation proposed so far—i.e., following rules as acting according to them—, then we must conclude that non-human animals, plants, and mere mechanisms are systems capable of mastering concepts. But according to Intellectualism, acting in accordance with rules is not sufficient to credit someone with mastering concepts. A subject can only be credited with mastering concepts if and only if she is capable of *following rules* in a proper sense, i.e., to carry out conceptual activities in accordance with rules *because* she recognizes and understands them as such—she grasps, so to speak, their normative force (Davidson 1982, Brandom 1994, 2010, McDowell 2009b). Thus, while some plankton, a thermometer, or a chimpanzee are systems capable of displaying complex behavioral patterns in accordance with specific rules, a subject masters concepts because she is capable of recognizing, explicitly or implicitly[9], the rules governing their activities, i.e., she is capable of capturing the normative nature of rules and does not merely act in accordance with them. Since apparently only human animals are capable of acting in virtue of rules *as such* (*qua* rules) and not merely in accordance with them, nothing but human animals can think conceptual thoughts (Brandom 1994, 2002b, Davidson 1982, 1997, McDowell 1994, 2009b, Rorty 1979).[10] This is, broadly speaking, the core of Intellectualism.

9. See for instance Brandom 1994, chap. 1.

10. According to Intellectualism, only human animals are capable of acting in accordance with rules as such (qua rules) since, apparently, solely they have the cognitive tool required for acting in that way: language. While non-human animals have sensitive abilities to act in accordance with laws, only humans have the *spontaneity* required to follow rules in the proper sense. But what does spontaneity have to do with rules and language? At least since Kant, spontaneity or freedom is considered as the fundamental condition for rule-following. As Kant pointed out: "… freedom, among all the ideas of speculative reason, is also the only one whose possibility we *know* a priori because it is the condition of the moral law, which we do know" (Kant 1788, 5/4). According to intellectualists, only human animals are free in the sense that they are the only ones capable of being constrained by their own laws. While non-human animals are constrained by external causal laws, only human animals are capable of being constrained by their own "interpretation" of laws (autonomy of the will). But humans are capable of constraining themselves (by their own laws) because—intellectualists argue—they have language, and language is—according

I have pointed out that thinking requires—in the intellectualist view—the ability to grasp and bring reasons supporting such thoughts into play in reasoning. But now I might be asked: what do reasons have to do with mastering concepts and following rules? It is widely argued by intellectualists that in order to think a thought, it is not enough to be sensitive or responsive to reasons. A mouse fleeing, for instance, can be perfectly credited as responding to a particular reason: danger. Similarly, a chunk of iron rusting can be credited as responding to a reason as well: the presence of moisture. Nevertheless, neither the mouse nor the chunk of iron is responding cognitively. What is required to respond in that way—intellectualists argue—is to "be sensitive to the normative force of reasons" (Brandom 2010, 14), i.e., to be responsive to what McDowell has called "reasons *as such*" (McDowell 2009b, 128). In effect, what makes a response a cognitive one is that it has been carried out not just for a good reason but because the subject has *recognized* (explicitly or implicitly) that there was a good reason for responding in that way.[11] Otherwise—intellectualists point out—there would be no grounds for holding that a subject was thinking (conceptually) instead of merely responding differently to stimuli (Brandom 2002b, 2010, McDowell 2009b). A three year old girl who utters the words "That's red" in front of a red object would hardly count as judging the object as being red if, when asked, she were unable to bring into play the reasons in virtue of which she responded as she did. In Brandom's terms, there would be no cognitive difference between the little girl and a parrot which has been trained to respond to red things by uttering the same noise ("That's red"). Neither of the two would count as making an authentic conceptual move within the realm of perceptual judgments (see Brandom 2002b).

At this point, the connection between grasping reasons and mastering concepts should be evident. After all, acting for reasons as such (qua reasons) and following rules involve exactly the same cognitive ability. Let me clarify this.

We know that in order to think a thought, mastering the relevant concepts involved in such a thought is required. But we also know that to

---

to them—the tool required to carry out the *evaluative-reflective* activity involved in rule-following (see Bermúdez 2010, Brandom 2002b, 2010, Davidson 1982, 1997, and McDowell 2009b).

11. Again, intellectualists do not require that in order to respond in a cognitive way, one must actually recognize (reflect about) whether the reasons are good ones. As McDowell clearly points out: "[a]cting for a reason, which one is responding to as such, does not require that one reflects about whether some consideration is a sufficient rational warrant for something it seems to recommend. It is enough that one could" (McDowell 2009b, 129).

master a concept F is not just having the ability to apply F according to a rule (the rule that prescribes whether something falls under F or not) but having the ability to apply F according to a rule *because*—and this is the core of the intellectualist proposal—one recognizes its normative force, i.e., one realizes that one *ought* to take a certain object as falling under F *because* one recognizes that the object satisfies the conditions prescribed by the rule. Under careful examination, it will be noticed that the word "because" does not lend itself to ambiguities. For, it reveals that following rules requires responsiveness to rules as such (i.e., to respond in accordance with rules qua rules). But rules and reasons are just two sides of the same coin. In effect, a mouse which acts in accordance with the rule that prescribes fleeing whenever one is presented with danger is responding to a particular reason: danger. Equally, a mouse which responds to danger by fleeing is acting in accordance with a specific rule: the rule that prescribes fleeing whenever one is presented with danger. Therefore, both to act for a reason as such (qua reason) and to respond in accordance with a rule as such (qua rule) involve the same cognitive ability. But since following a rule in the proper sense is to respond in accordance with a rule as such (qua rule), it must be concluded that to act in virtue of reasons as such (qua reasons) and to follow rules are exactly the same thing: each one implies the other.

An obvious consequence that follows from the intellectualist view is that crediting a subject as having a *contentful* thought implies crediting her as capable of acting for reasons as such (qua reasons). Let us see why. Content is usually said to be the way a state represents (is about) an object or event. Put in these terms, content is meant to be the *normative* property of a state. It provides its correctness conditions by prescribing the circumstances under which it is true or false, right or wrong, accurate or inaccurate. Now, we have seen that thinking thoughts implies having the ability to master the concepts involved in such thoughts. But according to Intellectualism, to master a concept F is just to grasp the norm that prescribes whether something falls under F or not, and to grasp a norm is just to follow a rule. Since the conceptual content of thoughts rests upon its conceptual components (concepts), it follows that the normative properties of thoughts (the possibility they may be right or wrong) rest upon rule-following. But since rule-following and acting for reasons as such (qua reasons) involve the same cognitive ability, it must be concluded that the normative properties of thoughts, their contents, rest upon the activity of reason-giving. Without being able to grasp and bring reasons into play

in reasoning—intellectualists argue—it is impossible to be credited with contentful thoughts at all.

2. *Traditional Empiricism and the myth of the Given*

Intellectualism is not a completely consistent tradition. At the moment, it is possible to identify at least two conflicting lines within the intellectualist approach: (a) a full-blooded holistic (or coherence) line and (b) an empiricist line. The full-blooded holistic line is related to rationalism in that it claims that nothing can count as a reason for holding a thought except another thought (Davidson 1982, 1983, Rorty 1979, Brandom 1994).[12] The empiricist line, on the contrary, is concerned with the idea that nothing but perceptual experiences can *ultimately* serve as reasons for thinking empirical thoughts (Brewer 1999, 2005, McDowell 1994, 2009a). According to this proposal, perceptual experiences must provide reasons in the sense that they must directly allow us access to the world in a cognitive way. Otherwise—empiricists argue—empirical thoughts would never be about the empirical world. In Kant's terms, they would be empty (cf. McDowell 1994, 4f.).

For the sake of my argumentation throughout this article, I will not focus on the holistic line. In what follows, I am going to focus on the predicament those with an empiricist spirit find themselves in when following the intellectualist approach. To begin with, I will argue that Traditional Empiricism collapses because it supports both the intellectualist approach to thought and a non-intellectualist approach to the cognitive role of perceptual experience. However, my main inquiry does not end here. In the third and last section, I will argue that McDowell's Minimal Empiricism collapses as well because of his attempt to defend an over-intellectualized and contradictory variety of empiricism, namely, a conceptual empiricism that preserves both the intellectualist approach to thought and a conceptual but *passive* and, consequently, a non-intellectualist approach to perceptual experience. Yet it would not be wise to anticipate conclusions: prudence is a virtue that will eventually pay off. Let us see why the traditional empiricist theory collapses.

12. As Davidson has put it, "[w]hat distinguishes a coherence theory is simply the claim that nothing can count as a reason for holding a belief except another belief" (Davidson 1983, 141).

Empiricism represents a powerful form of Intellectualism. To understand why this is so, let us focus on the empiricist proposal. Empiricism, says Bertrand Russell quoting the Encyclopedia Britannica, "is the theory that all knowledge is derived from sense experience" (Russell 1936, 131). According to this definition, empiricism can be characterized in two ways: on the one hand, as a theory of how empirical thoughts acquire empirical content and, on the other, as a conception of the foundations of our empirical knowledge (knowledge of the empirical world). Intuitively, empiricism is concerned with both content-acquisition and justification of our empirical thoughts. It is certainly true that empiricism is supposed to explain how, through perceptual experience, our thoughts acquire empirical content; but it is also true that empiricism is concerned with providing a foundation of empirical knowledge. Although it might be natural to disconnect these two dimensions (the content-acquisition dimension and the thought-foundation dimension), there is a conceptual dependency between them. Indeed, the idea that perceptual experiences must ultimately justify our empirical thoughts follows from the assumption that empirical thoughts acquire content by means of experience. This is so because if experience has the role of providing empirical thoughts with content (e.g., with the content "this is red"), then empirical thoughts must refer to (be about) what experience illuminates (e.g., a red object); but if empirical thoughts refer to what experience illuminates, it is then natural to appeal to experience to access (to recognize) the truth-conditions of thoughts.[13]

Now, the idea that the foundational role of perceptual experience rests upon a particular content-acquisition theory is not random. In fact, it is a consequence that follows the intellectualist approach underlying the whole empiricist theory. As mentioned before, thinking thoughts requires mastering the concepts involved in such thoughts. Consequently, thinking empirical thoughts involves mastering empirical concepts—i.e., concepts based on observations. But we already know that according to Intellectualism, having a concept requires the ability to follow a rule that prescribes whether or not something falls under such concept. Consequently, mastering empirical concepts (e.g., "red") involves the ability to follow *empirical rules*, i.e., rules that govern the appropriate use of empirical concepts. And here empiricists appear on the scene. For, according to them, only

13. It is intuitive to think that if experience is the source of empirical thoughts, then it must be also the source of their verification. Otherwise, what else but experience might judge if the understanding has acted in accordance with the instructions it was given?

perceptual experience is capable of providing what Wittgenstein (1953) has called an "interpretation" of empirical rules. Let me clarify this line of thought.

We know that, according to Intellectualism, acting in accordance with rules is not enough to follow rules in the proper sense. To be sure, following a rule involves acting in accordance with what the rule prescribes *because* one has recognized its normative force, i.e., one has recognized that one is presented with the circumstances prescribed by the rule and consequently acts in accordance with what the rule prescribes. Empiricists accept this intellectualist view. But just because they accept it, they bring perceptual experience into play. This is because, according to them, only perceptual experience can allow us directly into the empirical circumstances we must recognize in order to follow empirical rules in the proper sense.

To make my point more explicit, let us focus on the language theory of Logical Empiricism.[14] This theory represents one of the most significant attempts to explain how our empirical knowledge claims (i.e., claims that express knowledge about the empirical world) acquire *meaning* (empirical content). Guided by the idea that language is a system of symbols governed by rules, logical empiricists express the association between words and objects of experience, which is fundamental to explain the meaning of empirical knowledge claims, in terms of semantic rules that assert: "whenever you are presented with such and such objects (e.g., a red object) utter such and such words (e.g., 'red')" (cf. Ayer 1954, Schlick 1959). As a result, one grasps the meaning of empirical knowledge claims (they acquire empirical content) if and only if one grasps semantic rules; and, in principle, one grasps these rules if and only if one is capable to respond according to what the rules order (e.g., utter "red") in the circumstances prescribed by the rules (e.g., whenever one is presented with a red object).

Now, it is true that semantic rules order to perform certain linguistic actions (e.g., to utter "red") in front of certain observable circumstances (e.g., in front of red objects). However, their role is not limited to this. The interesting thing about these rules is that they not only order to perform such linguistic actions in the circumstances prescribed by the rules but also to *obey* them—i.e., to perform such linguistic actions *because* (and here the intellectualist ingredient comes to light) one has recognized one is

14. In the following, I will understand "Logical Empiricism" in the narrowest sense of the term, i.e., as referring only to the theory that asserts that the semantic properties of empirical knowledge claims rest upon perceptual episodes.

presented with the circumstances prescribed by the rules. As a result, one grasps a semantic rule if and only if one is capable of obeying it, i.e., of responding according to what the rules order (e.g., uttering "red") in the circumstances prescribed by the rules (e.g., whenever one is presented with a red object) *because* one has perceptually recognized one is presented with the circumstances prescribed by the rules (e.g., with red objects). Without such perceptual-recognitive capacity, without a perceptual *acquaintance* with the circumstances prescribed by the rules, the idea of someone as grasping semantic rules—and consequently as having empirical verbal knowledge—would make no sense. Our empirical knowledge claims—empiricists argue—would lack the constraints required in order to be about the empirical world (cf. Russell 1936, 133).[15]

Let us examine now the reason the theory of Logical Empiricism collapses. Semantic rules are introduced to explain the normative properties (meaning or content) of empirical knowledge claims. In effect, these rules are introduced to shed light upon the normative connection between words (concepts) and objects-properties of the empirical world. However, instead of explaining this connection, all rules do is just bring into play a more basic normative connection between a perceptual-recognitive mental state and the world. But what is wrong with appealing to perceptual-recognitive mental states? Recognizing that one is presented with the circumstances that a rule states is an *action*, namely, one that may be correct or incorrect by virtue of how things are in the world. But then it is reasonable to ask: where do the normative properties of these perceptual-recognitive states (the possibility that they may be correct or incorrect) derive from? Obviously, they cannot derive from semantic rules because these states

15. In "The Limits of Empiricism" (1936) Russell paradigmatically exemplifies this requirement. Russell distinguishes between (a) verbal knowledge that, properly speaking, can constitute knowledge and has the necessary logical properties required to serving as premises in an inference and (b) nonverbal sense-knowledge that, as such, is not properly speaking structured knowledge and, therefore, has no epistemic and logical properties, but, unlike mere sensitivity, it is a sort of awareness, identification, or recognition of objects and properties (see also Russell's distinction between "knowledge by description" and "knowledge by acquaintance" in Russell 1912). According to Russell, for a sentence "there is a cat" to be genuine empirical knowledge, she who proffers it must: (i) know a nonverbal sense-episode, (ii) acknowledge that she said "there is a cat", and (iii) realize or recognize that she said "there is a cat" *because* of (i), i.e., she must realize that she has expressed "there's a cat" *because* she is in a state of nonverbal sense-knowledge about (i.e., she perceptually recognizes) a cat. According to the interpretation I am supporting, this Russellian argument makes explicit the intellectualist intuition that underlies any form of empiricism, namely: we cannot think empirical thoughts (or our thoughts cannot acquire empirical content) unless we are able to recognize in perception the circumstances that make them true.

are introduced in order to explain the normative properties of knowledge claims. Consequently, logical empiricists need to introduce *new* rules into the perceptual realm in order to explain the normative properties of such perceptual-recognitive states. But this leads to a vicious regress. For, in order to recognize that one is presented with the circumstances prescribed by these new rules, it is required to appeal to some new re-recognitive mental states. But in that case, where do the normative properties of such re-recognitive mental states derive from? The vicious regress is inevitable. In order to avoid such vicious regress, however, logical empiricists have no option but to just take for *granted* the normative connection between these perceptual-recognitive states and the world. And this is exactly what logical empiricists do. But then they are faced with "givenness in its most straightforward form" (Sellars 1991, 167). For, they illegitimately appeal to episodes of perceptual awareness whose own normative properties are not only explained but simply *taken as given*, and, therefore, they are unable to explain what they were really introduced to explain: the normative properties of linguistic episodes.[16]

It is important to note that semantic rules are introduced to explain not only the normative properties of knowledge claims but also their epistemic authority. In fact, logical empiricists have been primarily concerned with the thought that semantic rules would be a suitable tool to preserve the foundational dimension of perceptual experience. Briefly, the argument asserts the following: when one *grasps* an empirical knowledge claim, one understands that the claim is true, because to grasp empirical knowledge claims means to follow semantic rules, and to follow semantic rules means to claim such knowledge claims *because* one has recognized the circumstances prescribed by the rules—i.e. the circumstances that make them true. Then, empirical knowledge-claims have authority just because they are the result of following semantic rules (cf. Schlick 1959).[17]

It goes without saying that "signs" of the myth transpire in such explanation. Empirical knowledge claims are conceived as actions whose authority derives from the fact that they follow semantic rules. But in order to be

16. I want to thank José Giromini for bringing this powerful interpretation of the Sellarsian concept of "givenness" to my attention. It should be noted that this interpretation is closely related to some Sellarsian ideas developed in "Some Reflections on Language Games" (see Sellars 1954).

17. It is important to notice that the logical empiricist theory (in particular Schlick's theory) concerning the authority of empirical knowledge claims (or observation reports) makes explicit what I have suggested before regarding the conceptual connection between the "foundational" and the "content-acquisition" dimensions of empiricism.

credited as properly following a semantic rule, it is required to be able to perceptually recognize the circumstances prescribed by the rule. And this is where logical empiricists struggle with the same problem. For, to recognize the circumstances is an action that can be correct or incorrect. But then where does the authority of such actions derive from? Obviously, it cannot derive from the rules of language, because these actions are introduced in order to explain the authority of such linguistic episodes. Consequently, logical empiricists are faced again with two options: either they introduce new rules in the perceptual-recognitive realm to explain the authority of these recognitive states in a vicious regress, or they just take their authority for granted. And this is exactly what they do. But then they are faced again with "givenness in its most straightforward form" (Sellars 1991, 167). For this time they illegitimately appeal to "self-authenticating" episodes of perceptual awareness whose own authority is not only explained but simply *taken as given*, and therefore they are unable to explain what they were introduced to explain: the authority of linguistic episodes. As Sellars clearly pointed out in the famous section VIII ("Does empirical knowledge have a foundation?") of *Empiricism and the Philosophy of Mind*:

> … *if* observation reports are construed as *actions*, *if* their correctness is interpreted as the correctness of an *action*, and *if* the authority of an observation report is construed as the fact that making it is "following a rule" in the proper sense of this phrase, *then* we are face to face with givenness in its most straightforward form. For these stipulations commit one to the idea that the authority of *Konstatierungen* rests on nonverbal episodes of awareness—awareness *that* something is the case, e.g. *that this is green*—which nonverbal episodes have an intrinsic authority (they are, so to speak, "self-authenticating") which the *verbal* performances (the *Konstatierungen*) properly performed "express" (Sellars 1991, 167).

At this point, we should not be surprised at the conclusion drawn by Sellars. For, the myth of the Given is just a consequence or symptom that follows from a widespread disease derived from bringing together two conflicting theses: (a) the intellectualist thesis, according to which the normative properties of cognitive states that have empirical content (whether they be public claims or private thoughts) depend (rest) upon the activity of reason-giving—i.e., activity that essentially involves rule-following—, and (b) the intuitive empiricist thesis, according to which the normative properties of cognitive states that have empirical content derive (*ultimately*) from perceptual experience. This is so because, if we accepted both, we

would be forced to conclude, on pain of vicious infinite regress, that, unlike thoughts, perceptual experiences grant us direct cognitive access—and, consequently, not by means of rule-following—to the empirical aspects of the world. But then, we would be faced with "givenness in its most straightforward form". For, we would be appealing to some sort of *sui generis* episodes of perceptual awareness whose own normative properties are not explained but simply taken as given. In brief, what I am suggesting is that the very concept of "the Given" emerges from bringing together an intellectualist approach to thought and a cognitive but non-intellectualist (i.e., not resting upon rule-following) approach to perceptual experience. Avoiding one of them, the givenness immediately disappears.[18]

3. *Minimal Empiricism and McDowell's intellectualist predicament*

In this section, I will focus on the new Minimal Empiricism defended by John McDowell (1994, 2009a). This new form of empiricism does not seem to be indifferent to Traditional Empiricism. As I have already argued, part of what it is to be an empiricist is inherently related to being an intellectualist. In fact, we have already seen that being an empiricist involves assuming both (a) the intellectualist approach to thought and (b) a non-intellectualist approach to perceptual experience. In the previous section, I have pointed out that to the extent to which traditional empiricist theory supports (a) and (b) simultaneously, it collapses. In this section I shall argue that, endorsing the same two approaches, McDowell's Minimal Empiricism collapses as well.

It is quite well-known that McDowell wants to rescue the empiricist view without falling into the myth of the Given. Roughly speaking, his strategy consists in providing experience with the normative properties of thoughts in order to avoid the problem traditional empiricists are faced with: givenness. Specifically, McDowell argues that experience is conceptual. McDowell seemingly supposes that by attributing it a con-

18. One possible alternative is just preserving the intellectualist approach but dismissing the empiricist view—i.e., the idea according to which the normative properties of empirical thoughts derive from perceptual experience. This strategy is defended by proponents of the full-blooded holistic line (see for instance Rorty 1979, Davidson 1984, and Brandom 1994). Another possible alternative is just defending a cognitive-epistemic approach to perceptual experience but dismissing the intellectualist approach. Currently, this strategy is defended by some non-conceptualists which argue for an externalist point of view regarding the justificatory relations between experiences and thoughts (see for example Alston 2002 and Burge 2003).

ceptual nature, perceptual experience might serve as a *tribunal* of thoughts constraining their contents in a rational way without falling into the emptiness of episodes whose own normative properties are merely taken for granted.

It is important to make a comment on this new form of conceptual empiricism. Supporting the idea that experience is conceptual, McDowell draws upon the Kantian distinction between intuitions and concepts, or more precisely, between receptivity and spontaneity, so as to argue that an accurate understanding of perceptual experience must conceive it as the result of spontaneity (the understanding) already at work on receptivity (sensibility), i.e., as the result of concepts already exercised on intuitions. Experience—McDowell argues—is a state or event that, though essentially sensible, has already been conceptualized or thought-out by the understanding (McDowell 1994, lecture 1).[19]

Accepting this suggestion, the following issue arises: spontaneity is the realm of reasons, i.e., the realm of what is truly normative. In this realm, one applies concepts in virtue of rule-following. Consequently, in the spontaneity, the understanding is free to think thoughts, i.e., it is free to follow rules. We have already seen that in the empiricist view the understanding must be rationally constrained in order to think thoughts about the world, and this includes, of course, the ability to think empirical thoughts—i.e., thoughts based on observations. But the understanding cannot provide to itself such constraints on pain of emptiness.[20] Accordingly, empiricists need to introduce experience—and this is something McDowell agrees with—as the cognitive capacity required to provide the understanding with such constrictions. Broadly speaking, experience is the faculty suitable for presenting us with the circumstances under which empirical thoughts are true and which we have to recognize in order to make sense of empirical thoughts as referring to the empirical world (McDowell 1994, xii).

We have already noticed that an appropriate way of interpreting what is required by the intellectualist account is through rules. This means

19. In McDowell's terms: "[t]he relevant conceptual capacities are drawn on in receptivity […] It is not that they are exercised on an extra-conceptual deliverance of receptivity. We should understand what Kant calls 'intuition'—experiential intake—not as a bare getting of an extra-conceptual Given, but as a kind of occurrence or state that already has conceptual content" (McDowell 1994, 9).

20. If the understanding gave to itself the rules that prescribe the circumstances that make thoughts true, then thoughts would be trapped under the will of the understanding and, therefore, could never be about anything but themselves. As Kant has clearly put it and McDowell suggests: "Thoughts without content are empty" (cf. Kant 1787, B75, and McDowell 1994, 3).

that in order to think correctly about objects and events of the world, following *empirical rules* is required, and to follow these rules means to think (or to judge) what the rules order in the circumstances prescribed by the rules because one has recognized one is presented with such circumstances. Consequently, McDowell has no alternative but to argue that subjects acquire thoughts about the empirical world because they are able to follow empirica rules, and they are able to follow such rules because they are able to recognize, through perceptual experience, the circumstances prescribed by the rules (the circumstances that make their thoughts true).

Now, we have already seen that to perceptually recognize is an action, and to be precise, one that may be correct or incorrect. But then where do the normative properties of such actions derive from? McDowell might respond: because of their conceptual nature, the normative properties of perceptual recognitions derive from the rules of the understanding, i.e., the rules that prescribe the appropriate application of concepts (and this includes empirical concepts). But if this answer is taken into careful consideration, McDowell heads toward a non-virtuous *circular* explanation, for he introduces perceptual experience in order to explain the normative properties of thoughts, yet when he is asked to provide an explanation of the normative properties of those perceptual states, he appeals to the rules he wants to clarify through perceptual experience: the rules of conceptual thoughts.[21] In order to avoid such circularity, McDowell is faced with the two well-known options logical empiricists have faced before: either McDowell introduces *new* rules into the realm of perceptual experience which are different from the understanding, leading to a vicious regress, or he takes their normative properties for granted. Both alternatives, we know, are unsatisfactory.

21. McDowell could argue that there is no problem with such circularity (see for instance McDowell 1994, Postscript to Lecture III). For, perceptual experience belongs to the realm of reasons and, consequently, it is governed by the rules of the understanding—i.e., the rules that prescribe the appropriate application of concepts in thoughts. Nevertheless, arguing this, McDowell would be contradictory with his empiricist spirit according to which the normative properties of empirical thoughts derive from (rest upon) perceptual experience, because if the normative properties of perceptual experience belonged purely to the realm of the understanding (the realm of conceptual thoughts), then there would *not* be any normative *priority* between perceptual states and thoughts. All of them (perceptions and thoughts) would rest upon the same rules. But then, McDowell's Minimal Empiricism would turn into a full-blooded holistic or coherence theory and, consequently, he would be faced with the same problem that, according to McDowell (see for example McDowell 1994, Introduction), coherence theories are faced with: the normative disconnection between mind and world.

Of course, McDowell might avoid this predicament just by making a simple movement: perceptual experience—he might argue—is not the result of an activity and, therefore, should not be recognized in terms of actions. Although experience is conceptual, the concepts in experience are *passive*.[22] This means that, in experience, concepts are not exercised—i.e., are not applied by rules. While in the realm of thoughts one judges that such and such is the case, that is, one applies concepts by rule-following, in experience, on the contrary, one is simply invited to judge that such and such is the case—i.e., one is free to apply concepts. Consequently, experience does not give rise to the problem of explaining the normative properties of these episodes, because since they are not the result of a *free-concept-application* activity, their normative status does not arise from the fact that they follow rules properly.

Although attractive, this strategy collapses under its own weight. We already know that, forced by the intellectualist approach, McDowell is committed to the idea that thoughts acquire empirical content if and only if they are thought in virtue of reasons *qua* reasons—i.e., in virtue of rule-following. Consequently, in order to think about the empirical world (e.g., about red objects) it is required to be able to recognize that one is presented with the circumstances prescribed by empirical rules (e.g., with red objects). Let us now turn to the problem. If perceptual experience illuminates, invites, or presents us with the way the world is (e.g., it presents us with such and such object as being red), but it is not yet an active part of our cognitive abilities, then it does not serve the purposes for which it is introduced. Because, with the introduction of such passive experience, all McDowell has done is simply change the circumstances one needs to recognize in order to think about the world, namely: from the circumstances of the world to the circumstances that experience is "inviting" us to capture. But now, we must *actively* capture (i.e., recognize) these circumstances just as they are presented to us in experience.

Let me clarify this point. Suppose, for example, that our experience invites us (or presents us) with a content of the form "this is red". Although our experience sets before us such conceptual content, in order to think

22. This is something McDowell effectively argues. In effect, according to McDowell: "… when we enjoy experience conceptual capacities are drawn on in receptivity, not exercised on some supposedly prior deliverances of receptivity. And it is not that I want to say they are exercised on something else. It sounds off key in this connection to speak of exercising conceptual capacities at all. That would suit an activity, whereas experience is passive. In experience one finds oneself saddled with content" (McDowell 1994, 10).

"this is red" we still have the task of applying concepts actively—for, although in experience concepts are passive, in thoughts and judgments they are exercised actively. In this case, we are to judge that the concept "red" is appropriate to capture in thought the conceptual content that experience is inviting us to capture. But then, the same problem arises. For, according to the intellectualist view, thinking (or judging) that something is red is conceived in terms of an action governed by rules that order to perform such actions (e.g., to apply the concept "red") whenever one is presented with the appropriate circumstances (e.g., red objects). Consequently, if someone is credited with thinking that something is red, she must also be credited with being capable of recognizing she is presented, not now with a red object, but with an experience of the form "this is red"—since according to McDowell, perceptual experience presents us directly with the empirical world. The sequence is well-known. Recognition is understood as an action that can be correct or incorrect, in this case, by virtue of how the world is presented to us in experience. But then, where do the normative properties of such actions derive from? We know it is impossible to appeal to the rules of the understanding on pain of circularity. Consequently, McDowell is faced again with the same two unacceptable options: either he introduces new rules in the perceptual-recognitive realm in order to explain the normative properties of these actions, leading to a vicious regress, or he just takes falling into the emptiness of "givenness" for granted.

What I am suggesting is that if perceptual thoughts, which are conceptual in the sense they are thought of in virtue of rules, are based on perceptual experiences, and if these perceptual experiences, in turn, are conceptual in the sense that concepts guarantee their normative properties, then an open question arises regarding how we must interpret those conceptual capacities in perception. If such conceptual capacities in perception are active in the sense conceptual capacities in thoughts are, then the explanation either becomes circular or falls into the myth of the Given. But if those conceptual capacities are to be understood as passive capacities, then McDowell's Minimal Empiricism collapses. For, according to Intellectualism, being normative means being governed by rules qua rules, and being governed by rules qua rules essentially involves a cognitive activity, namely: the activity of responding to rules as such.

In a nutshell, the problem McDowell's empiricism is faced with does not arise because of the conceptual interpretation of experience. What makes it collapse is the irreconcilable no-win situation in simultaneously

maintaining both the intellectualist approach to thought and the empiricist intuition according to which perceptual experiences are the ultimate normative *source* of empirical thoughts. By taking "passivity" into account, McDowell sets aside perceptual experience from the demands required by the intellectual approach. But by doing this, its normative status collapses. With passivity, it is certainly possible to avoid the "givenness", but simply because passivity destroys the normative properties of any mental state. With activity, in contrast, perceptual experiences may regain the normative properties in the sense required by the intellectualist approach, yet at the expense of leading McDowell's empiricism to circular explanations or to "givenness". McDowell seeks to have his cake and eat it: on the one hand, he wants experience to be a passive state, i.e., not governed by the rules of the understanding, but, on the other, he wishes experience to be a normative episode, i.e., one governed by such rules. In other words, McDowell wants experience to be some kind of passive presentation or appearance *before* the mind, as well as some kind of active mental *awareness*—i.e., to be *part of* the mind. McDowell's Minimal Empiricism collapses under its own weight.

4. *Conclusion*

I have argued that Intellectualism is incompatible with any form of empiricism. First, I have argued that Traditional Empiricism collapses because it brings together two conflicting theses: the intellectualist thesis, according to which the normative properties of thoughts depend (rest) upon the activity of reason-giving, and the intuitive empiricist thesis, according to which the normative properties of empirical thoughts derive from perceptual experiences. Second, I have argued that McDowell's Minimal Empiricism collapses as well because of his attempt to make sense of an over-intellectualized and contradictory variety of empiricism: one that preserves both the intellectualist approach to thought and a conceptual but passive and, consequently, a non-intellectualist approach to perceptual experience.

References

Alston, William 2002: "Sellars and the 'Myth of the Given'". *Philosophy and Phenomenological Research* 65, 69–86.

Ayer, Alfred 1954: "Basic Propositions". In: *Philosophical Essays*. London: MacMillan Press Ltd, 105–124.

Bermúdez, José Luis 2010: "Two Arguments for the Language-Dependence of Conceptual Thought". In: Julia Langkau & Christian Nimtz (eds.), *New Perspectives on Concepts* (= *Grazer Philosophische Studien* 81). Amsterdam: Rodopi, 37–54.

Brandom, Robert 1994: *Making It Explicit: Reasoning, Representing and Discursive Commitment*. Cambridge, Mass.: Harvard University Press.

— 2002a: "Holism and Idealism in Hegel's Phenomenology". In: *Tales of Mighty Dead: Historical Essays in the Metaphysics of Intentionality*. Cambridge, Mass.: Harvard University Press, 178–209.

— 2002b: "The Centrality of Sellars's Two-Ply Account of Observation to the Arguments of 'Empiricism and the Philosophy of Mind'". In: *Tales of Mighty Dead: Historical Essays in the Metaphysics of Intentionality*. Cambridge, Mass.: Harvard University Press, 348–367.

— 2010: "Conceptual Content and Discursive Practice". In: Julia Langkau & Christian Nimtz (eds.), *New Perspectives on Concepts* (= *Grazer Philosophische Studien* 81). Amsterdam: Rodopi, 13–35.

Brewer, Bill 1999: *Perception and Reason*. Oxford: Clarendon Press.

— 2005: "Perceptual Experience Has Conceptual Content". In: Matthias Steup & Ernest Sosa (eds.), *Contemporary Debates in Epistemology*. Malden: Blackwell, 217–230.

Burge, Tyler 2003: "Perceptual Entitlement". *Philosophy and Phenomenological Research* 68, 503–548.

Davidson, Donald 1982: "Rational Animals". In: *Subjective, Intersubjective, Objective*. Oxford: Oxford Clarendon Press, 95–105.

— 1983: "A Coherence Theory of Truth and Knowledge". In: *Subjective, Intersubjective, Objective*. Oxford: Oxford Clarendon Press, 137–157.

— 1997: "The Emergence of Thought". In: *Subjective, Intersubjective, Objective*. Oxford: Oxford Clarendon Press, 124–134.

Evans, Gareth 1982: *The Varieties of Reference*. Oxford: Clarendon Press.

Kant, Immanuel 1788 (2002): *Critique of Practical Reason.* Indianapolis: Hackett Publishing Company.

McDowell, John 1994: *Mind and World*. Cambridge, Mass.: Harvard University Press.

McDowell, John 2009a: "Experiencing the World". In: *The Engaged Intellect: Philosophical Essays*. Cambridge, Mass.: Harvard University Press, 243–256.
— 2009b: "Conceptual Capacities in Perception". In: *Having the World in View: Essays on Kant, Hegel, and Sellars*. Cambridge, Mass.: Harvard University Press, 127–144.
— 2009c: "Hegel's Idealism as Radicalization of Kant". In: *Having the World in View: Essays on Kant, Hegel, and Sellars*. Cambridge, Mass.: Harvard University Press, 69–89.
— 2009d: "Self-Determining Subjectivity and External Constraint". In: *Having the World in View: Essays on Kant, Hegel, and Sellars*. Cambridge, Mass.: Harvard University Press, 90–107.
— 2009e: "Why is Sellars's Essay Called '*Empiricism* and the Philosophy of Mind'?". In: *Having the World in View: Essays on Kant, Hegel, and Sellars*. Cambridge, Mass.: Harvard University Press, 221–238.
Rorty, Richard 1979: *Philosophy and the Mirror of Nature*. Princeton: Princeton University Press.
Russell, Bertrand 1912: "Knowledge by Acquaintance and Knowledge by Description". In: *The Problems of Philosophy*. Oxford: Oxford University Press, 46–59.
— 1936: "The Limits of Empiricism". *Proceedings of the Aristotelian Society* 36, 131–150.
Schlick, Moritz 1959: "The Foundation of Knowledge". In: Alfred Ayer (ed.), *Logical Positivism*. New York: Free Press, 209–227.
Sellars, Wilfrid 1954: "Some Reflections on Language Games". *Philosophy of Science* 21, 204–228.
— 1991: "Empiricism and the Philosophy of Mind". In: *Science, Perception and Reality*. Atascadero, CA: Ridgeview Publishing Co, 127–196.
Wittgenstein, Ludwig 1953: *Philosophical Investigations*. Oxford: Basil Blackwell.

Grazer Philosophische Studien
90 (2014), 253–272.

PLEASURES OF THE COMMUNICATIVE CONCEPTION

Uku TOOMING
University of Tartu

Summary
In this paper, I criticize Christopher Gauker's approach to the attributions of desire which identifies them with commands on behalf of others. These are supposed to be needed in situations wherein such commands have to be qualified in some way. I argue that his account doesn't manage to make explicit the need for the concept of desire, and I defend my alternative according to which desires are related to our understanding of how commands on a person's behalf relate to her subjective satisfaction.

1. *Introduction*

A large part of everyday life consists of thinking about what one wants and what to do about it. That people want things, are disposed to act on their desires, and are frustrated when they don't get what they want, seems to be a trivial, yet important, truth about us.

"What does it mean to want something?" is by no means a trivial question, however. This question can also be presented as follows: what must be so of a person for it to be acceptable to attribute a desire to her? Possible answers vary. Some would argue that to count as having a desire the person needs to token certain internal states and these states can be identified with the desire in question. Another view would be that it suffices for the subject to have certain functional relations instantiated between her inner states and her behaviour. But there could also be a view according to which, for instance, an acceptable attribution only requires certain behavioural criteria to apply but it doesn't require the attributee to have any specific internal states or functional roles. The question "what does it mean to want something?" can have different answers depending on how strong commitments are read into the conditions of acceptable attributions.

Views about desire attributions can also differ in what they require of the role of such attributions within the life of human beings and in the needs to which attributions are supposed to be primarily responsive. A standard thought is that their role is that of explaining and predicting behaviour. Given such a view, there is a quite natural, though not inevitable, pull towards the idea that attributers are committed to some internal states or the instantiation of functional roles in order to make the explanatory and predictive aims one has intelligible. Nonetheless, there is logical space for other approaches that conceptualize the need for desire attributions in a different manner and in which case the question about commitments might merit a different answer.

Christopher Gauker's conception of desires and desire attributions is one of those accounts which refuses to conceive of the primary function of these attributions in explanatory and predictive terms. It also belongs to a collection of views which acknowledge only minimal commitments with regard to desire attributions, excluding, for instance, a commitment even to behavioural dispositions. According to his communicative conception, as I interpret it, the need to ascribe desires arises in linguistic practice when people present commands on others' behalf in specific contexts. This means that by attributing desires we are making certain communicative moves towards one another. Also, to have a desire, nothing more is required than being the appropriate subject of such vicarious commands. The communicative conception, then, professes to give both an account of the attribution and the nature of desires.

In this paper, I will focus on the issue of attributions.[1] The aim is to see whether the communicative conception construes desire attributions in a way that lets us see the full significance of the concept of desire. I am going to argue that Gauker is unable to articulate a need for desire attributions which couldn't be satisfied by any other means. Because of this, he is also unable to explicate the conditions under which it is appropriate to ascribe a desire. His account thus blocks us from seeing the full commitments of those attributions. With that in mind, I will present an alternative account which doesn't have these problems but which is still to a large extent congenial to the communicative conception. It maintains the core idea that the ascriptions of desire involve commands on others' behalf.

1. There are different theories about the nature of desires available (Smith 1987, Stampe 1987, Strawson 1994, Oddie 2005, Schroeder 2004). Although I won't directly address these theories in this paper, I still hope that the present considerations have some consequences for deciding which of them is feasible and which isn't.

The plan is as follows. I start by articulating Gauker's account and discussing the motivation behind it. Then I turn to the aforementioned problems with his conception and argue that it has difficulty explaining the need to adopt the concept of desire as long as the latter is to have any substantial significance. After that I will present my alternative account, according to which desires are to be understood in terms of the connection between commands and the recognition of pleasure. I will argue that it fares better in articulating the need for desire attributions while still acknowledging their communicative significance.

2. *Desires and the communicative conception*

To understand Gauker's motivation for defending his conception of desires, we should briefly look at his theory of linguistic communication because his conception of attitudes is meant to support it. Gauker's general project is already explicitly laid out in his first book, *Thinking out Loud* (1994).[2] His target there is what he takes to be the standard model of communication, presumably going back to Locke, according to which the function of communication is to express one's thoughts to hearers (Gauker 1994, 3). Gauker's positive goal has been to develop an alternative conception of linguistic communication according to which it functions to help people achieve their collective goals (ibid. 4). This distinction figures within his latest book as the distinction between a conveyance conception and a cooperative conception of communication (Gauker 2011, 217f).

The difference between the two may not seem intuitively very substantial. One could see them as complementary because communication seems to serve both expressive and cooperative purposes. Gauker isn't willing to concede that. The difference becomes more salient when we take into account how Gauker has developed his cooperative conception. For instance, the goals of conversation are supposed to be determined by the objective context of the conversation. This kind of objective context consists of a set of sentences that the speakers should accept in order to achieve the cooperative goal of the conversation (Gauker 1998). The objective context, then, depends on the collective goal that the conversation has—if the subjects conformed to this context, then they would be able to achieve that goal (Gauker 2011, 219). Such a collective goal is presumably

2. Although one could go further back to find the sketch of this project (see Gauker 1992).

irreducible to individual desires. If that were not so, Gauker would face a threat of falling back to the conveyance conception. The interlocutors would then be required to express their individual attitudes first in order that the goals of communication be achieved, which is exactly what the conveyance conception claims. One crucial difference between the two accounts of communication, then, concerns the role that attitudes (such as beliefs and desires) play in explaining both the function and success of linguistic communication. According to the conveyance conception, these attitudes are explanatorily prior to the language that speakers use to convey them. According to the cooperative conception, on the other hand, linguistic communication is explained at least at its most fundamental level without any reference to the attitudes of interlocutors.

Gauker's communicative conception of beliefs and desires is motivated by his aim to attack the conveyance conception of linguistic communication and to defend his alternative. For Gauker one of the reasons for the appeal of the received view is the intuition that beliefs and desires are language-independent theoretical entites which are attributed to others in order to explain and predict their behaviour and which are identifiable with physical states in the brain. Gauker calls it the "postulationist" conception (Gauker 2003a, 216). Under the assumption of postulationism, it is natural to think that linguistic utterances function to express the attitudes of speakers, the nature of which is independent of language. This would amount to the conveyance conception of communication. It should be noted, though, that postulationism and the conveyance conception don't entail one another and that Gauker acknowledges this (ibid.). In what follows, I won't address the question of which theory of communication is the right one. I focus on the opposition between postulationism and Gauker's own view.

Unlike postulationists, Gauker doesn't think that attributing beliefs and desires commits the ascribers to the existence of causally efficacious inner states. Instead, he thinks that attitudes can be understood in terms of specific types of speech act. The attribution of them is supposed to be an extension of the practice of linguistic communication and not something that explains it as it was envisioned by the conveyance conception. If we now focus on desires, ascribing a desire to someone is to be equated with commanding on her behalf or in her stead (Gauker 2003a, 221).[3] So when

3. As for beliefs, Gauker understands them in terms of assertions on another's behalf. Since the focus of this paper is on desires, I leave evaluating the communicative conception of beliefs for another occasion.

we think and talk about others wanting something, we are not speculating about the causal underpinnings of their behaviour (as postulationists claim). We are instead simply making vicarious commands. Gauker himself illustrates the attribution of desires with the following hypothetical story. A house is being built. One of the builders, Balam, wants some rope and orders his assistant, Namu, to bring some. Namu can then tell the keeper of the supplies that Balam wants some rope. In doing that he simply passes over Balam's command (Gauker 2003a, 222). So the initial suggestion is that the need to attribute desires is the same as the need for the means of conveying the speech acts of other people. As for explanatory contexts wherein we seemingly ascribe desires to others in order to explain their behaviour, Gauker tries to interpret them as reconstructions of conversations on another's behalf (ibid., 252). For instance, if I say that John went to the store because he wanted to drink milk, we can reconstruct it as us commanding on John's behalf to drink milk. Gauker doesn't really deny, then, that the attribution of desires may serve explanatory purposes, but this is supposedly parasitic on the communicative role of attributions.

It is noticeable that the story about Balam and Namu doesn't really reveal any discursive need over and above the need to speak about commands. It really doesn't add anything to the linguistic practice besides a new term which simply re-labels the commands and claims that are already in use. Nonetheless, there are resources in Gauker's theory that enable us to adjust the aforementioned story in a way that implies a more distinctive role for the concept of desire.[4] This role comes from the fact that, according to the communicative conception, commands on another's behalf form a specific kind of discursive move that is responsive to a need which can't be satisfied by mere commands. There are situations wherein mental state ascriptions have an application that can't be fulfilled by the mere ascription of a speech act. These situations include, for instance, those wherein the attributee doesn't perform the corresponding speech act herself or wherein the attributer isn't ready to take the attributee as authoritative (Gauker 1994, 273). Let's call these two cases the occasion of *absent utterance* and *declined authority* respectively. With this distinction in mind, the adjustments to the original story are relatively simple. We have to imagine basically the same community as was depicted by Gauker. However, we focus upon situations wherein the agents find it appropriate

4. This doesn't mean, though, that concepts for Gauker are something over and above linguistic items (Gauker 2011). In the framework of the communicative conception, one can talk about terms and concepts interchangeably.

to command on another's behalf without acknowledging their authority or wherein they haven't produced the corresponding utterance themselves. Having the concept of desire, then, fulfils the need to perform vicarious speech acts in such situations.

We should address another worry. It is probably quite counter-intuitive to equate the attributions of desire to X with commands on X's behalf. Commands are made to an audience but one can have desires that are not addressed to others. If a person wants there to be peace in the world, does it make sense to say that it is appropriate to command on her behalf that there be peace? Who is this command directed at? Yet one can amend Gauker's conception by replacing commands with a more general category of evaluative utterances. Such utterances have the form "It should be the case that *p*" (cf. Van Cleave & Gauker 2010, 320). If we adopt this idea, the attribution of a desire that *p* should be taken as a claim on other's behalf that it should be the case that *p*. Here, I am still using the term "command" to mark such an utterance, but it is important to bear in mind that the term takes both the second- and third-person form.

So what kinds of ontological commitment does the communicative conception of desires have? To answer this question we should ask where exactly in the philosophical landscape it is located. Gauker opposes postulationism, but exactly what kind of philosophical position can be identified with the postulationist conception? Functionalism is probably Gauker's primary target because he also characterizes postulationism as a position according to which beliefs and desires are inner mechanisms that mediate between sensory inputs and behavioural outputs (ibid., 215), and it is functionalism that is known for analysing mental states in terms of their causal roles. But the communicative conception also seems to clash with identity theory, which equates having mental states (either types or tokens) with having certain physical structures in one's head. All in all, the communicative conception is opposed to any view that tries to find a deeper fact of the matter which the desires are supposed to be dependent upon or identical with. In that sense, it has strong affinities with those strands in the philosophy of mind that are labelled "interpretationism" (cf. Child 1994, 23), and in his earlier book Gauker also adopts this designation. He defines it as follows: "an account of what beliefs are need not be anything over and above an account of their attribution" (1994, 293).

What distinguishes the communicative conception from other interpretationist positions is the way each depicts the function of folk psychological interpretation. As I've already mentioned, Gauker denies that the use of

desire attributions in the explanation and prediction of behaviour is the most fundamental form that their ascription takes and instead stresses their role in linguistic communication. This brings him into conflict with Dennett, for instance, for whom the most notable value of having our folk psychological vocabulary is tied to the prediction of behaviour (Dennett 1987, 17). The communicative conception also tends to be in tension with Davidson's views, according to which attitudes are causally efficacious and identical with physical events (Davidson 1980). In sum, the communicative conception is a version of interpretationism. It equates someone having desires with her being interpretable in a certain way. The most fundamental form that this interpretation takes is identified with making claims on another's behalf in specific situations and not with attributing causally efficacious inner states to them.

Is there any motivation for accepting the communicative conception of desires apart from Gauker's wish to defend a particular conception of linguistic communication, according to which the function of linguistic utterances is explained independently of individual attitudes? I think there is. First, the communicative conception takes the attributers of desires to make minimal commitments concerning what psychological facts need to hold for the attributions to be acceptable. Attributers only make commitments to existence of whatever explains the ability to utter and comprehend speech acts. This should be of interest to those who think that mentalistic talk that ordinary people use is to a large extent impervious to psychological findings. If the communicative conception is feasible, then there's a reason to take seriously such a view of folk psychology's autonomy from scientific psychology.[5] Second, desire attributions do have communicative significance—it is open to the attributer whether to endorse or disapprove of them. This means that Gauker's account presents us with the possibility of understanding the function of desire attributions primarily in terms of such a communicative significance, while the explanatory and predictive role remains only secondary. If the communicative conception is correct, it presents an original take on the primary need for desire attributions—the latter consist primarily of treating others as (virtual) interlocutors not of attempts to explain and predict their behaviour.[6] I take it, then, that

5. This, of course, relates to the issue of distinguishing between the subpersonal and personal levels of mental processes (Dennett 1969, Hornsby 2000) and the putative susceptibility of folk psychology to scientific findings (Churchland 1981).

6. There are other accounts of the function of folk psychology that oppose the centrality of explanatory and predictive roles. Zawidzki (2008), for instance, has suggested that desire attri-

the communicative conception of desires can be of interest regardless of whether Gauker's model of communication is accurate.

Nevertheless, I will argue that the communicative conception doesn't manage to explain why the concept of desire should be introduced to the conceptual repertoire and that there's a more substantial explanation available.

3. *Problems with the communicative conception*

3.1 *The critique*

We can bring into focus what is lacking in the communicative conception if we focus on the question what role the concept of desire fulfils. In section II we defined the cases of *absent utterance* and *declined authority.* These are situations wherein talking straightforwardly of commands on another's behalf is felt to be somewhat inappropriate and where introducing the concept of desire seems to be useful for drawing the distinction between full-blown and qualified vicarious claims. We can now ask whether this move by Gauker suffices to make explicit the motivation for introducing the concept of desire. I will argue that the communicative conception in its present form still doesn't manage to articulate a substantial conceptual need behind the concept of desire. To see this, we should first consider what the notion of "conceptual need" actually amounts to.

A conceptual need isn't strictly speaking a psychological state. It is rather a state of affairs which is defined in terms of the absence of the concept in question. It arises when people find their conceptual repertoire insufficient for conveying something distinctive about the world or themselves. It can be appealed to when we are interested in understanding the main benefits that the application of the concept offers. If a proposed conceptual need can actually be fulfilled by means which are already available, then we have a reason to doubt that such a proposal is acceptable. It is acceptable only when the proposed need really motivates the introduction of the concept that we want to explain. One can find a similar appeal to conceptual needs in the genealogical approach to concepts pursued by such authors as Bernard Williams (2002) and Edward Craig (1990). Such approaches

butions have a distinctive mindshaping role; Andrews (2012), on the other hand, has proposed a pluralistic view of folk psychology. Gauker's account can be taken as yet another alternative.

present us with a hypothetical situation wherein the concept isn't yet in use and then ask what needs could be fulfilled by introducing an expression which corresponds to that concept into the vocabulary of the community. This kind of speculation is intended to bring out the distinctive role of the concept that is under scrutiny and its connection to other concepts. The present question, then, is whether Gauker has been able to do that with the concept of desire.

So what are those conditions which necessitate the introduction of a new concept? The point of introducing a concept varies in a way that depends upon the nature of the thing or the property it is meant to track. Gauker himself distinguishes different ways of thinking about and explaining the nature of a thing: 1) characterizing its internal structure; 2) its function; 3) its place in human conventions (Gauker 2003a, 272). It seems to me that the third way actually forms a subset of the second because if we individuate a property by its place in human conventions (such as married status or citizenship), we are still making a functional claim. The communicative conception can be taken as the claim that by attributing desires we identify a certain communicative act which counts as such due to convention. This would situate desires both in the second and in the third type.

I understand convention to be a set of assumptions governing some social practice. Under what conditions is a new term required to label an element in such a practice? This brings us to a rather difficult issue because it isn't even clear if practices can be individuated without mentioning the mental states of the community members. Setting that worry aside, it is reasonable to claim at least that a minor change in a convention doesn't mean that the practice it pertains to or its particular elements need be renamed. A social category can remain constant through minor changes as long as it fulfils its function. For instance, after the presence of a priest has been deemed unnecessary, marriage can still be called "marriage", the bridegroom can still be called "bridegroom", and bride be called "bride". The conceptual needs for introducing terms for social practices and statuses are quite various, but the introduction of a new term should have distinctive discursive and practical consequences. In talking about distinctive consequences I am relying on intuitive considerations, but the following discussion should demonstrate how they can be put to work in evaluating Gauker's account.

Think now of agents who have come up with the practice of commanding on each other's behalf. Let us again consider those occasions when the

ascription of desires, instead of commands, seems to become necessary. There were at least two kinds of situation wherein the straightforward command on another's behalf could be deemed inappropriate—when the other person doesn't make the corresponding command herself (absent utterance) and when she is not taken as authoritative (declined authority). In a nutshell, the question is: should the agents who engage in making claims on another's behalf in such situations adopt a new concept to distinguish these vicarious utterances from those that are made when the relevant utterance is not absent and authority is transmitted?

First, we should note that there's a problem with the suggestion that declined authority *and* absent utterances necessitate introducing a new concept. Assuming that both situations are cases of genuine conceptual need, the question arises: given that these two types of cases are very different, why should they motivate the introduction of the same concept? It seems that if they both present genuine conceptual needs, they should give rise to two concepts. One would identify non-authoritative commands on another's behalf. The other would label commands on other's behalf without the latter having made the relevant utterances. These two concepts seem to diverge in their significance. The situations of declined utterance and absent authority, then, seem like competing cases for introducing the concept of desire.

So let's consider the two cases separately, starting with declined authority. Does this situation articulate the need to adopt the concept of desire? I am not sure if the answer has to be affirmative. The case of declined authority doesn't seem to create a genuine conceptual need at all because the fact that the attributer isn't sure about the attributee's authority doesn't change the fact that the latter still performed the speech act that the subject now performs on her behalf, so it would be unnecessary to invent a new label for this act. It would be sufficient to say: "She commanded that *p* but don't take her seriously."

Note that by denying that the case of declined authority reveals a real conceptual need, I don't base my claim merely on the fact that the application of a new term in a sentence can be paraphrased into a sentence that expresses the attribution of a command with a qualification. It is rather about asking what point there is to the introduction of a new term. Of course it is useful to distinguish between authoritative and non-authoritative commands. Nonetheless, it seems to be actually less confusing to make this distinction by maintaining the term "command" in both cases than by adopting a new term. If we maintain the term, we won't lose sight of the

fact that the respective utterance was also uttered in the case of declined authority. For instance, if a child tells her parent to bring her ice cream, it should be unproblematic to still report it as a command with a comment that it was merely a child who commanded it.

It is possible to think up a situation wherein introducing a new term in this case would have a more far-reaching significance. I don't deny that. For instance, one can imagine a community where those who make commands in the case of declined authority are sanctioned or even punished. In such a case the introduction of a new concept would have distinctive practical consequences. But imagining such a possibility requires us to make very particular (and rather far-fetched) assumptions about the community in question. It is problematic to invoke such a conceptual need to explain the concept that is in use in our actual world. I take it, then, that the case of declined authority doesn't motivate the introduction of the concept of desire. There should be a more substantial and stable benefit that it offers.

As for the case of absent utterance, why do we need to speak about desires in a situation wherein we are inclined to command something on another's behalf but wherein they haven't commanded the same thing themselves? Why not simply say that they *would command* it if the circumstances were more accommodating? Why does one need to coin a new term for such a discursive situation? Doing it doesn't seem to answer a real conceptual need but simply invents a label for something that can already be expressed in terms of (potential) commands. When a new concept is introduced to represent a form of social practice, the latter has to be genuinely new with distinctive practical consequences. If the concept of desire merely labelled the dispositions to command, it wouldn't satisfy any genuine conceptual need. So the initial response to this proposal is the same as in the case of declined authority.

To counter this problem, Gauker might stress that the ascription of a desire to someone doesn't even require that she is disposed to produce the corresponding utterance (Gauker 2003a, 225). But if the attributee need not even be disposed to make the corresponding command that is uttered on her behalf, then an obvious question arises about the grounds for performing the vicarious speech act. This shouldn't be an entirely arbitrary matter, but Gauker himself doesn't provide any suggestions about the conditions under which the attribution would remain appropriate without the disposition on the attributee's part. Yet without these grounds, the vicarious commands seem to be ill-motivated. What's more, this proposal leaves us in the dark about what could be the need to make such seem-

ingly baseless commands on another's behalf without those others even being disposed to utter those commands themselves. In order to make it explicit, one has to say something more about the conditions under which such vicarious commands can be made. Otherwise this practice (and the introduction of a concept to mark it) wouldn't make much sense. There should be a reason to make such a command.

Gauker might respond by making some further distinctions between different kinds of vicarious commands in order to account for the intuition that the attribution of real desires should say something about the grounds for making vicarious commands in the case of an absent utterance. For example, he could distinguish between person-relative and non-person-relative commands. The first would imply that the person on whose behalf it is made endorses it, but the latter would leave it open. The first case would perhaps provide us a condition under which the command on other's behalf is appropriate, even if the other hasn't performed the corresponding speech act herself. What's more, wouldn't the need to make vicarious person-relative commands also suffice to account for the need to introduce a new concept? After all, it would allow the agents to relate commands on behalf of others to the subjective perspective of the attributees. It makes explicit that the latter are ready to endorse such commands. This seems like a practically significant and distinctive move in the discourse.

This suggestion can be related to my own positive proposal that will be laid out in the fourth section, but it involves an ambiguity as it stands. The person-relative command on another's behalf either requires that the person explicitly endorses it or that she would endorse it in certain circumstances. In the first case, the need for introducing a new concept loses its urgency—one could simply attribute to her the claim that it should be the case that *p*—after all, she explicitly endorses it. But if she merely *would* endorse it, then one can always ask about the grounds on which she would do that. The answer to that question should say something about the person, but the communicative conception leaves us in the dark about what it is. In any case, what the endorsement of the vicarious command amounts to needs further elaboration, and Gauker hasn't done that. My own positive proposal in the fifth section will exploit this lacuna in his account.

My claim, then, is that the communicative conception doesn't manage to explain why one would need to coin new terms in the practice of performing vicarious speech acts, thus arriving at the idea of desires. It either gives the concept a communicative role which is too shallow to motivate introducing a new term or it faces the problem of explaining the grounds

for making vicarious claims, unable as it is to make the need for the new concept of desire intelligible. This still isn't a conclusive proof against the communicative conception. One could maintain, for instance, that the concept of desire really does play such a shallow role. But if an alternative explanation which demonstrates a more substantial conceptual need and which makes explicit the grounds for vicarious commands were available, it would have an advantage over Gauker's position.

3.2 *The solution?*

The inability of the communicative conception to explain the need for the concept of desire poses a challenge. How should we proceed in order to remedy the situation? I suggest that we can mostly retain Gauker's main idea that the attribution of desires involves making claims on behalf of others, but we need to complement it by reconsidering how it relates to the conceptual needs that the agents might have in the case of absent utterance.

The explanation of the concept of desire that I have in mind starts with the following hypothetical situation. The agents have linguistic competence and the ability to make commands. In that respect, it follows Gauker's lead. It also acknowledges that those agents command on another's behalf and that sometimes this may happen in the cases of absent utterance and declined authority. But I claim that the explanation of why a concept of desire is needed in such a community should also make explicit why it is proper to command on other's behalf even if the other didn't make the corresponding utterance herself. You need to have a reason to claim something on another's behalf. This means that there should be an explanation of why this kind of action is appropriate. It is natural to say that someone's having a desire *explains* why it is appropriate to make a corresponding command on her behalf. The question about the grounds of vicarious claims, then, should be the key to revealing the conceptual need for the concept of desire. It is not merely a nuisance that needs to be faced after the communicative conception has already been established.

Is this concession a return to the postulationism which Gauker was opposing? It certainly differs from his account in admitting that the concept of desire has a distinctively explanatory role. But since I haven't yet answered what kind of an explanatory posit it is, one can leave open whether desires are causal postulates or not. I can also continue to agree with Gauker that the attribution of desires doesn't have much predictive potential. But the problems I've presented mean that one cannot stay as

deflationist as Gauker does. The account of desires needs to say something about the conditions under which a person is properly describable in terms of such a concept. I don't exclude the possibility that commands on behalf of others could form the core of the concept of desire, but this is only a starting point for a more substantial account which I am going to provide in the next section.

4. *Modified communicative conception*

Let us, then, reconsider the concept of desire. I hope that it is a relatively intuitive point that one can distinguish between mere acknowledgements of commands and actual attributions of desires. The latter should also take into account the perspective of the attributee. From the previous analysis we saw that the most promising account of the conceptual need—that of person-relative commands—connected desire attributions with the subjective perspective of the attributees. It remained obscure what this exactly amounted to. My proposal is that the ascription of a desire opens up the question of whether the interpreted person will also be satisfied when the represented state of affairs is realised. The ascription of a command doesn't have such a consequence. By saying that another person orders or would order something one doesn't necessarily say anything about her actual sympathies. Gauker is, of course, more subtle in that the practice of making commands on behalf of others is meant to be separable from the actual utterances of individuals. But as we saw in connection with the problem of grounds, Gauker doesn't provide many positive suggestions concerning the ascribability conditions of such vicarious commands. The alternative account should make explicit how desire attributions explain when it is appropriate to command on another's behalf and relate this somehow to the attributer's understanding of the attributee's subjective satisfaction.

To see how to do that let us now reconsider the hypothetical situation wherein the concept of desire isn't yet introduced, but wherein the agents engage in the practice of making commands. It is reasonable to assume that such agents have also some primitive abilities for social cognition such as emotion-recognition and understanding how people are affected by events: whether they are pleased or displeased, satisfied or frustrated. This claim is also plausible for empirical reasons (see Nichols (2001, 436) for the view that the understanding of affect doesn't require possessing the ability to attribute propositional attitudes). One quite obvious move

to make, if our aim is to clarify what subjective satisfaction amounts to, is to take into consideration our capacity to recognize when someone is pleased.

Doesn't the admission of affective understanding already bring with it the understanding of desires? After all, it is quite natural to say that if someone finds something pleasurable, she wants it. One can doubt, though, that the mere fact of something being disposed to cause pleasure to a subject amounts to her having a desire for it. What is pleasurable for a subject need not be desirable for her. What's more, from the present perspective, introducing a new term "desire" for something that can already be satisfactorily understood in terms of potential pleasure is pragmatically rather pointless.

Another important thing to note is that the understanding of pleasure doesn't imply that the agents in question have to be postulationists about it. They don't need to identify pleasure with some causally efficacious inner states. In order to understand that another person is pleased with a state of affairs, nothing more is needed than certain observable criteria that help one recognize that the other is in that condition. These criteria can include linguistic utterances, bodily expressions and temporally extended patterns of behaviour—anything that is relevant for recognizing that the person is satisfied. The present appeal to the understanding of affect, then, isn't opposed to the spirit of the communicative conception.

Now the question is what the concept of desire could enable the agents in such a hypothetial situation to do. They understand evaluative utterances and are able to realise when a person feels pain or pleasure, is content or frustrated. In this kind of state, circumstances will emerge in which a person who utters a command doesn't seem to be personally satisfied when it is carried into effect. There are also situations wherein she presumably takes pleasure in some state of affairs but hasn't commanded it. The need to speak about desires arises when there is a tension between the objective satisfaction condition of a command and the subjective affect that it arouses. These are situations wherein the person's commands don't seem to reflect her actual needs. The ascription of a desire is essentially perspective-sensitive because it not only states a command on another's behalf but also relates it to the question of what really pleases that individual.[7] In this way desires are still akin to commands, but they are also tied to the

7. It is left open whether subjective satisfaction is something directly given and introspectively accessible to the subject or not. But this is a further question that I won't discuss here.

perspective of the agent, the understanding of which is lacking in Gauker's communicative conception. The conceptual need that the notion of desire satisfies is, then, the need to know which commands would lead a person to her subjective satisfaction and which would not.[8] In addition, this proposal nicely explains the grounds for making vicarious commands in cases of absent utterance by relating them to the ascribers' ability to anticipate and recognize others' pleasure. Gauker lacked the means to make explicit what the grounds for vicarious commands in such cases could be. The present account, on the other hand, provides a reason to treat the other person as if she had uttered a command because satisfying the latter would presumably be pleasing to her.

It is also noticeable that this line of thought forces us to recognize the explanatory role of the concept of desire. The modified communicative conception brings together Gauker's idea of the attribution of a desire as a command on behalf of others and the idea that the mental state concepts play an explanatory role. But it is important to note that, as with admitting the understanding of affect into our account, this concession doesn't imply postulationism about desire attributions. The latter would be the case if the explanatory role of desires consisted of tracking causally efficacious inner states. In the present case the explanation merely makes intelligible why the vicarious utterance is appropriate. As already noted, the recognition of someone being pleased doesn't require postulating inner causes behind a person's expressive behaviour. The desire attribution only requires for its success that the attributee would be satisfied if the claim on her behalf were fulfilled. That can be understood on the basis of overt criteria.

One might wonder whether my proposal encounters a similar difficulty as Gauker's initial account did, namely, whether the need to understand which commands would please the attributee is really such a substantial conceptual need for the concept of desire. It could be argued that introducing a new term into such a situation is unnecessary. Why not simply speak about dispositions to feel pleasure in certain situations? The key here is that vicarious claims still play a role in the present account. Since my own proposal retains Gauker's idea that desire attribution involves commands on another's behalf, the concept of desire can't be reduced to that of a disposition to feel pleasure. It articulates a connection between speech acts and affect that helps its users achieve something distinctive in the linguistic practice. Note that this claim about distinctiveness doesn't appeal

8. This doesn't imply that satisfaction of a desire actually always leads to pleasure.

to the impossibility of paraphrasing the desire-ascriptions into vicarious commands conjoined with pleasure-attributions. The paraphrase might be possible just like it was possible to talk about qualified command-reports instead of desires in the case of declined authority. The distinctiveness in question concerns what desire attributions enable us to do. They have a synthetic function of articulating the connection between two kinds of phenomena: imperative speech acts and affective conditions. What's more, they explain why it is proper to extend the practice of commanding to the cases of absent utterance. The original communicative conception didn't extend the use of the concept beyond the domain of speech acts and didn't assign any explanatory role to the concept. The present proposal, on the other hand, does both.

Let's elaborate on the merits of the present account. Explaining the concept of desire in terms of pleasure also connects it naturally with its wider functional significance in human social life, thus demonstrating the explanatory potential of the modified communicative conception. Having the concept of desire as a tool for connecting the communicative behaviour of agents with their affective expressions enables people to think about whether making claims on another's behalf will actually be advantageous to her when the command gets fulfilled. This opens up specific ways of coordinating with others. The best way to show this is to consider what the cooperative actions would look like if people didn't attribute desires to one another. Such actions would still have goals in relation to which the behaviour of individuals could be evaluated. The imagined community could still be engaged in building houses and shelters, gathering and growing foodstuff, educating children, performing rituals, etc. During these activities people would command each other to do things and also to command on behalf of one another. All of this fits into Gauker's account alongside the fact that the mere practice of vicarious commands doesn't require the introduction of the concept of desire into the communicative practice. What that imagined community lacks is the means of connecting the question of whether the practice of commanding is successful or not and the question of whether participants of that practice will be subjectively satisfied with the particular claims made.

This is basically a repetition of my earlier points. But these points are important to bear in mind if we are to recognize the full significance of having the concept of desire. Introducing that concept into the aforementioned situation enables people to understand the success of their communicative practice in a new light. Now the success can be evaluated

not only on the basis of the fulfilment of objective goals but also in terms of individual agents' satisfaction with it. Taking the idea that people have desires seriously might even cancel the objective criteria for evaluating the success of the practice. At least the desire-based criteria might compete with the objective ones. Assuming that people are somewhat altruistically motivated, at least to their ingroup members, the need to know whether the communicative practices of a group bring satisfaction to its members is important. Employing the concept of desire enables us to satisfy that need. In that sense this concept is significant for leading a worthy communal life and for coordinating action.

But does the concept of desire also open up new ways of manipulating others? I don't deny that at all. For example, one can deceive others into thinking that a person wants some state of affairs to be realised even though the attribution actually isn't sensitive to the affective profile of that person but instead only to that of the subject/attributer. Let's consider a situation wherein another person is more authoritative in her wishes than the attributer. In that case the latter might achieve what she wants by putting others to work on the false assumption that the person whose vicarious commands they are satisfying actually wants them to be satisfied. However, in actual fact, they are serving the subject who has attributed such a desire to the more authoritative person. By exploiting the fact that desire attributions are supposed to bring together the commands on another's behalf and the other person being pleased, one can deceive people into mistakenly seeing such a connection. The present example is only one among many strategies to do that. It is almost needless to say that these manipulative moves undercut the original purpose of the desire attributions. If people always decoupled vicarious commands and subjective satisfaction from one another, then the practice of attributing desires would presumably lose its point.

The modified communicative conception, then, handles the intuition that desires involve subjective satisfaction very well. It makes explicit the functional significance of the concept of desire and offers an explanation for the cases of absent utterance. In all these respects it is superior to Gauker's original account. Nevertheless, it maintains that the attribution of desire involves commands on someone's behalf. It also avoids the postulationist claim that the attribution involves reference to the inner causes of people's behaviour. The addition of pleasure-recognition makes the account perhaps less deflationist, but that is the price we have to pay if we want to hold on to the main insights of the communicative conception.

5. *Conclusion*

The aim of this paper was to analyse the communicative conception of desires according to which the attribution of desire can be identified with the making of an evaluative utterance that something should be the case on other's behalf. Although it required substantial modifications, the main idea still remains intact. The modified communicative conception presents a feasible account of desire attributions that avoids postulationist commitments without having to deny that the attributions say something about the subjective perspective of the attributees. What this implies about the nature of desires is left for another occasion.[9]

References

Andrews, Kristin 2012: *Do Apes Read Minds? Toward a New Folk Psychology*. Cambridge, Mass.: MIT Press.

Child, William 1994: *Causality, Interpretation, and the Mind*. Oxford: Oxford University Press.

Churchland, Paul M. 1981: "Eliminative Materialism and the Propositional Attitudes". *The Journal of Philosophy* 78, 67–90.

Craig, Edward 1990: *Knowledge and the State of Nature. An Essay in Conceptual Synthesis*. Oxford: Clarendon Press.

Davidson, Donald 1980: "Mental Events". In his: *Essays on Actions and Events*. Oxford: Oxford University Press, 207–225.

Dennett, Daniel C. 1969: *Content and Consciousness*. New York: Humanities Press.

— 1987: *The Intentional Stance*. Cambridge, Mass.: MIT Press.

Gauker, Christopher 1992: "The Lockean Theory of Communication". *Noûs* 26, 303–324.

— 1994: *Thinking Out Loud. An Essay on the Relation between Thought and Language*. Princeton: Princeton University Press.

— 1998: "What Is a Context of Utterance?". *Philosophical Studies* 9, 149–172.

— 2003a: *Words without Meaning*. Cambridge, Mass.: MIT Press.

— 2003b: "Attitudes without Psychology". *Facta Philosophica* 5, 239–256.

— 2011: *Words and Images. An Essay on the Origin of Ideas*. Oxford: Oxford University Press.

9. Research in this paper was supported by Estonian Science Foundation Grant ETF9117.

Hornsby, Jennifer 2000: "Personal and Sub-personal: A Defence of Dennett's Early Distinction". *Philosophical Explorations* 3, 6–24.
Nichols, Shaun 2001: "Mindreading and the Cognitive Architecture Underlying Altruistic Motivation". *Mind & Language* 16, 425–455.
Oddie, Graham 2005: *Value, Reality, and Desire*. New York: Oxford University Press.
Schroeder, Timothy 2004: *Three Faces of Desire*. New York: Oxford University Press.
Smith, Michael 1987: "The Humean Theory of Motivation". *Mind* 96, 36–61.
Stampe, Dennis 1987: "The Authority of Desire". *Philosophical Review* 96, 335–381.
Strawson, Galen 1994: *Mental Reality*. Cambridge, Mass.: MIT Press.
Van Cleave, Matthew & Gauker, Christopher 2010: "Linguistic Practice and False-Belief Tasks". *Mind & Language* 25, 298–328.
Williams, Bernard 2002: *Truth and Truthfulness. An Essay in Genealogy*. Princeton: Princeton University Press.
Zawidzki, Tadeusz 2008: "The Function of Folk Psychology: Mind Reading or Mind Shaping?" *Philosophical Explorations* 11, 193–210.

Grazer Philosophische Studien
90 (2014), 273–288.

TELEOSEMANTICS, SWAMPMAN, AND STRONG REPRESENTATIONALISM

Uwe PETERS
King's College London

Summary
Teleosemantics explains mental representation in terms of biological function and selection history. One of the main objections to the account is the so-called 'Swampman argument' (Davidson 1987), which holds that there could be a creature with mental representation even though it lacks a selection history. A number of teleosemanticists reject the argument by emphasising that it depends on assuming a creature that is fictitious and hence irrelevant for teleosemantics because the theory is only concerned with representations in real-world organisms (Millikan 1996, Neander 1996, 2006, Papineau 2001, 2006). I contend that this strategy doesn't succeed. I offer an argument that captures the spirit of the original Swampman objection but relies only on organisms found in the actual world. The argument undermines the just mentioned response to the Swampman objection, and furthermore leads to a particular challenge to strong representationalist theories of consciousness that endorse teleosemantics such as, e.g., Dretske's (1995) and Tye's (1995, 2000) accounts. On these theories, the causal efficacy of consciousness in actual creatures will be undermined.

Introduction

Mental representations exhibit intentionality; they are about things or states of affairs. The things and states of affairs that they are about are their contents. For example, your mental representation of a dog wagging its tail is about a dog wagging its tail; it has a dog wagging its tail as its content. How is it possible for a mental representation to be about something?

A number of theories of intentionality have been proposed.[1] Arguably the most promising among them is teleosemantics (Millikan 1984, 2000, Papineau 1987, 1993, Dretske 1981, 1988). Teleosemanticists explain the

1. See Adams and Aizawa (2010), and Shea (2013) for an overview.

content of a representation *R* in terms of *R*'s biological function, where this function lies in the way *R* contributes to the biological end of the system using it for behavior guidance.

Teleosemanticists typically specify biological function in historical-etiological terms[2] as

> the upshot of prior processes of selection. A trait has a function if it has been designed by some process of selection to produce some effect. […] An effect of a trait counts as its function if the trait has a certain history: in the past possession of that trait produced the relevant effect, which in turn had the consequence [of] facilitating the reproduction of items with that trait. (Macdonald and Papineau 2006, 10f.)

According to teleosemantics, then, a state *R* in creature *C* will represent, say, snakes if *R* has the biologically designed function to be about snakes. And it has that function if it was in the past selected for registering snakes and initiating behavior advantageous in the presence of snakes.

The selection in question needn't always occur diachronically, over an evolutionary time span, but could take place via learning or conditioning synchronically, during the lifetime of an organism (Campbell 1974, Papineau 1984, Dretske 1988). Independently of whether it is selected for diachronically or synchronically, on the teleosemantic account, a state needs to have one of the two kinds of selection history in order to qualify as a representation.

One major objection to teleosemantics pertains specifically to the theory's commitment to selection history. The objection takes the form of the so-called 'Swampman argument', which, by asking us to imagine a creature that lacks any selection history, aims to show that beings without such history could arguably still have states with representational content (Davidson 1987, Braddon-Mitchell and Jackson 1997).

In response, "[m]ost proponents of teleosemantics" hold that the objection hinges on the assumption of a merely fictional creature, and since that is so "reject the idea that we should care about the Swampman intuition. It would be enough, they claim, if we could find a theory of referential

2. The historical-etiological view is not the only way in which teleosemanticists have understood biological function. For an alternative proposal see, for instance, Cummins (1975, 2002). Nonetheless, the historical-etiological view is advocated "by most teleosemanticists" (McDonald and Papineau 2006, 9). In this paper, only the majority view is at issue, and in what follows 'teleosemantics' should be read as referring specifically to the historical-etiological version of the theory.

content that was successful for real creatures" (Neander 2006, 385; see also Millikan 1996, Neander 1996, Papineau 2001, 2006). Call this the 'fiction response' to the Swampman objection.

In the following, I contend that this response fails. I provide an argument that captures the basic idea underlying the original Swampman objection but just assumes real creatures. As it turns out, this will not only dissolve the fiction response to the objection but also lead to a particular challenge to strong representationalist theories of consciousness that rely on teleosemantics, for instance, Fred Dretske's (1995) and Michael Tye's (1995, 2000) accounts. On these theories, the causal efficacy of consciousness in actual creatures will be undermined.

Before going into the details of the discussion, I begin with a brief recap of the Swampman argument.

I. *The Swampman argument and the fiction response*

According to teleosemantics, a state has the representational content it does in virtue of its biological function, and any state *X* has a biological function "*Y* if and only if *X* is now present because previous versions of *X* were selected in virtue of doing *Y*" (Papineau 1998, 1). Thus, as noted, for the teleosemanticist, for an organism to have states with representational content, these states are required to have a selection history.

Donald Davidson (1987) proposed the following thought experiment to challenge this view. Suppose Davidson is taking a walk in a swamp when he is suddenly struck by lightning. Suppose further that as soon as the lightning bolt has evaporated him, by random fluke, a perfect molecule-for-molecule replica of him reassembles itself out of the materials available in the swamp. Suppose finally that the replica of Davidson, call him 'Swampman', is behaviourally identical to Davidson; he walks and talks like him, greets his friends, writes philosophy papers, etc.

By assumption, Swampman will lack any state with a history of natural selection. Since that is so, according to teleosemantics, Swampman won't have any states with representational content. That is, he won't have beliefs, desires, intentions, etc. even though he is behaviorally entirely indistinguishable from a normal human being, i.e., Davidson, who *does* have beliefs, desires, intentions, etc. Critics of teleosemantics hold that this is a highly counterintuitive upshot of the theory and conclude that if teleosemantics has the consequence of denying Swampman

representational states, then it can't be an adequate account of mental representation.

There are at least two different strategies of responding to the Swampman objection available to the teleosemanticist (Neander 2012). The first is to attempt to ease the grip of the intuition that Swampman has representational states. The second is to grant the intuition but hold that it doesn't suffice to falsify teleosemantics.

In line with the second strategy, many advocates of teleosemantics propose what I called above the fiction response. They accept that denying Swampman mental states is counterintuitive but then maintain that this doesn't speak against their theory, for Swampman is imaginary and teleosemantics is only intended to be an account of mental representation in real creatures (Neander 1996, 124f; 2006, 385; Millikan 1996, 115f; Papineau 2001, 284; 2006, 185).

One particular way of spelling this response out is due to David Papineau (2001, 2006). For Papineau, arguing that teleosemantics is wrong because Swampman would *seem* to have states with representational content but no selection history is like arguing that water isn't H_2O just because one can imagine a possible world in which a different substance, say, XYZ plays the water role. As long as Swampmen "remain merely imaginary, they are no more relevant to teleosemantics than imaginary molecular make-ups are relevant to chemistry", Papineau (2006, 185) holds. He grants, however, that "actual" Swampman cases would "provide concrete evidence that teleosemantics is false" (Ibid), and thus

> present a real threat. True, a limited number of actual cases can sometimes be accommodated. A few actual examples of non-H_2O stuffs playing the watery role, rare molecules of heavy water (HDO), say, can perhaps be dismissed in the interests of overall theoretical unity or simplicity. ('We used mistakenly to think that was water, but now we know better.') But note that this move involves a real overriding of pre-theoretical usage, an alteration of what we say about actual cases, and this shift needs some substantial justification, in terms of increased simplicity or unity.
>
> Relatedly, if the counter-examples were frequent enough, and their dismissal couldn't be so substantially justified, then this would simply mean that the proposed reduction was false, and that the 'watery role', or the 'belief' and 'desire role', wasn't in fact filled by H_2O, or selectional states, after all. (Papineau 2001, 284)

II. *The argument revisited*

If the Swampman objection involves a merely imaginary creature, Papineau et al.'s fiction response is a natural move to make for the teleosemanticist. However, as I shall argue in this section, the objection can be reformulated in terms of actual organisms. This undermines the response and, in the absence of any other compelling reply to the Swampman worry, commits the teleosemanticist to the unattractive claim that there are actual creatures that lack representation even though they are behaviourally identical to conspecifics that do possess representational states.

In a first approximation to the point, let's agree that we have representational states and that we evolved from more primitive creatures. If that is so, then at some point in evolutionary history, representation must have emerged in the actual world.

Suppose, then, at some point in the past when organisms in this world haven't yet evolved representational states, there is a population of primitive creatures. One of them, call her '*C1*', acquires by random genetic mutation for the first time in evolutionary history a particular inner state *R*. As it happens, *R* is activated by and systematically co-varies with the presence of some object or state of affairs *X* in *C1*'s immediate environment and leads *C1* to exhibit behavior in response to *X* that has survival-promoting effects. As a result, *C1*'s life expectancy increases allowing her to transmit *R* to numerous offspring.

It is a common view among philosophers working on representational content that when a mental state systematically causally co-varies with some environmental condition, then it indicates or represents the latter (Stampe 1977, Dretske 1981, Fodor 1990). Given this view, since *R* in *C1* does systematically causally co-vary with *X*, one might propose that *R* represents *X* in *C1*.

There are well-known problems with this proposal, however. If systematic co-variance between *R* and *X* were sufficient, then *R* wouldn't only represent *X* but also various *X*-look-alikes, for the latter would, due to their resemblance with *X*s, have to manage to token *R* as well. If *X*-look-alikes were not able to cause *R*, then clearly *X*s wouldn't be able to do so either, for *X* is evidently a look-alike of itself. On the causal co-variance account, *R* would then have the content *X or Y or Z* etc. (where *Y*, *Z* etc. are *X*-look-alikes). However, representations don't have such disjunctive contents (Fodor 1990, 63ff). Worse still, since *R* would be about anything that it is tokened by, it could on the account at issue never misrepresent. And

since no state represents anything unless it can misrepresent, *R* couldn't be a representation after all (see Dretske 1981, chapter 8).

Teleosemantics proposes one way[3] of avoiding these problems by introducing the notions of biological function and selection history. On the teleosemantic account, *R*'s content is not specified in terms of *R*'s typical cause but rather its *effects*.[4] *R* is about *X*s and not *X*-look-alike non-*X*s just in case *R*'s being caused by the former rather than the latter had evolutionarily advantageous consequences for the organism with *R* and was in previous generations selected for producing those effects. So according to teleosemantics, *R* is about *X* if *R* was in the past selected for initiating behavior advantageous specifically in the presence of *X*, and not in the presence of *X*-look-alike non-*X*s. The truth condition of *R* is thus specifically *X*, and correspondingly *R* will misrepresent the environment when it is tokened by something else.

Returning with this to *C1*, the teleosemanticist will insist that, since *C1* didn't acquire *R* via inheritance from her ancestors but rather by random genetic mutation from one generation to the next, *R* in *C1* doesn't have a selection history and thus can't have representational content.

If that is so, however, then when exactly does representational content enter the picture according to teleosemantics? Consider, for instance, *C1*'s offspring. Since, by assumption, *C1*'s offspring inherit her novel capacity, behaviorally, *C1* and her offspring will be indistinguishable when they encounter *X*s. Furthermore, in *C1*'s offspring, *R* will also have a selection history, albeit a very short one, for it is just one generation old. As a result, on the teleosemantic account, in *C1*'s offspring, *R* should have some sort of representational content, for in *C1*'s offspring, *R* doesn't only do everything it does in *C1*, it also satisfies the historical-etiological condition that the teleosemanticist imposes on states with representational content. So *R* in *C1*'s offspring would be a representational state.

But then, since *C1* acts in the same way as her offspring when they detect *X*s, she seems to be on a par with Davidson's Swampman in the following respect: even though *C1* is, just as Swampman, behaviorally

3. The argument against teleosemantics that I outline in this paper suggests that another way of solving the problems of disjunctive content and misrepresentation is needed. One interesting proposal can be found in Bickhard (1993, 2004).

4. Unlike causal theories, which are input-based, the teleosemantic account is hence *output based*: Whether or not *R* systematically causally co-varies with *X*s or non-*X*s is irrelevant for its being about *X*s. What matters is that *R*'s registering *X*s and initiating a particular kind of behavior served the biological end of the consumer of *R* and was selected for doing so.

identical to a creature with representation, i.e., *C1*'s offspring, following teleosemantics, she would still lack any state with representational content. Using the familiar 'Swampman' terminology, *C1* would be just another 'Swamp-creature' for the teleosemanticist.

There are of course various differences between *C1* and Swampman. For instance, Swampman is by assumption molecule-for-molecule identical to a creature to which the teleosemanticist would grant representational content. In contrast, *C1* is not physically identical to her offspring in this sense. Furthermore, *C1* has ancestors and thus at least some kind of selection history, whereas Swampman lacks it entirely. But these differences aren't relevant here. The different physical constitution of *C1* and her offspring doesn't matter for the present argument because they lead in both *C1* and her offspring to the same causal-dispositional results.[5] They lead them to exhibit the same evolutionarily beneficial behavior when they detect *X*s. Also, even though *C1* has an evolutionary history, by assumption, her novel trait doesn't have such history. And it is only the representational/non-representational status of that trait that is at issue here. Thus, the same argumentative logic as in the original Swampman objection applies in the scenario introduced.

Note, however, that there is a difference between Swampman and *C1* that is crucial for present purposes. Since '*C1'* in the above scenario is just a placeholder for the first creature with any kind of representational content in the evolution of representation in the *actual* world, the Swampman argument can now be rephrased in terms of that actual creature. As a result, Papineau et al.'s fiction response to the original objection is undermined.[6]

One might object that a creature such as *C1* is about as imaginary, and thus irrelevant for teleosemantics, as Swampman. For *R* in *C1* is thought to attain representational content suddenly from one generation to the next, but since Darwin it is widely accepted that traits emerge rather *gradually* over many generations. Hence, *C1*'s acquisition of a representational state might seem as fictional as Swampman.

However, in the process of incremental changes that led from non-representational states to representational ones, representation must at some point have emerged. Perhaps the first representational state only differed

5. The worry about physical type identity can also be dealt with by rephrasing the argument, using an individual that undergoes selection in its lifetime. That individual prior to selection (via learning) will be physically identical to the individual after learning.

6. The reasoning here is similar to Macdonald's (1989). Thanks to David Papineau for bringing Macdonald's paper to my attention.

minutely from the non-representational state from which it arose. But a minute difference is all that is required for the argument above to get off the ground. For the variable '*R*' in the above scenario should be taken to refer precisely to the first, arguably, very unsophisticated representational state that might only slightly have differed from its non-representational predecessors. Since even on the gradualist picture, there must have been such a state at some point, gradualism about the evolution of representation in the actual world doesn't undermine the argument above.

If the argument can't be dismissed by holding that *C1* is an imaginary creature, however, then teleosemanticists are now committed to denying actual creatures such as *C1* representational content even though they are behaviorally equivalent to other actual creatures that *do* possess representational content. Note that *C1* isn't just a single outlier that could perhaps be ignored for the sake of greater theoretical unity. As a matter of fact, for any particular type of mental representation, there must have been a creature that, just like *C1*, came to be the first organism in the actual world with that representation.[7] Consequently, there were (and will be) plenty of Swampman-like creatures in the actual world. If the existence of such creatures "presents", as Papineau (2001, 284) holds, "a real threat" to teleosemantics, then teleosemantics does now face a real threat.

III. *From representation to consciousness and its efficacy*

While the preceding reasoning undermines the fiction response to the Swampman objection, there might be other replies to the objection that can equally well be applied to my revised argument. For instance, some teleosemanticists have reacted to Davidson's thought experiment by biting the bullet and rejecting the intuition that Swampman has mental states (see, e.g., Millikan 1996, Neander 1996). The same move could also be adopted in reply to the argument just introduced.

However, there are a number of problems with this response.[8] The one that I wish to highlight in the remainder of this paper becomes especially

7. Thanks to Janiv Paulsberg here for the generalization point.

8. For instance, Macdonald (1989) holds that if states such as *R* in *C1* (he considers "random mutants' proto-beliefs" instead) don't represent anything, it becomes hard to see how the selection of representation could get started in the first place (see also Cummins 1996, 46). Since *R* in *C1* doesn't have a selection history, it doesn't have a function, and hence can't represent. But if it can't represent, then *R*'s representing evidently can't be what has evolutionarily advantageous effects,

pressing when teleosemantics is part of one's theory of consciousness and one takes representational content to be constitutive of a conscious state. To see the problem at issue, a few words on accounts of consciousness are in order.

There are different theories of consciousness available. One particularly popular approach is representationalism (see, for instance, Dretske 1993, 1995, Tye 1995, 2000, Lycan 1996, Byrne 2001, Chalmers 2004). The theory explains what it is for a mental state to be phenomenally conscious in terms of the state's representing the world as being a certain way, that is, in terms of its having representational content.

There are weak and strong versions of the view. According to weak representationalism, conscious experience supervenes on representational content so that necessarily any two states that are the same with respect to the relevant representational content are the same phenomenally (Byrne 2001, McLaughlin 2003). The converse needn't be the case, however. In contrast, strong representationalism claims that conscious experience or phenomenal character is *identical* to representational content that meets certain further conditions (Tye 1995, 2000, Dretske 1993, 1995, Lycan 1996).

In what follows, I want to focus only on strong representationalism, that is, on the view that representational content that meets certain further conditions is constitutive of the conscious experience. The 'further conditions' phrase refers to the point that, since there are unconscious representations, for example, unconscious beliefs, or sub-personal representational states such those involved in early vision, more needs to be said about what makes representation constitutive of conscious experience.

One way of doing so is to hold that a representation *R* can only be conscious iff *R* has the right sort of content, and fulfills the right sort of functional role. For instance, for Tye (2000), the 'right sort' of content is (i) *abstract*, in that no particular concrete objects or surfaces enter into it, (ii) *non-conceptual,* in that the subject doesn't need to have the concepts required for specifying the content, and (iii) *intentional,* in that it doesn't

---

and thus can't become selected for. What will be selected for is *R* as a *non-representational* state. A non-representational state will then obtain a selection history but not a representation. Since selection doesn't add anything to the traits it operates on but only accounts for their propagation, it becomes difficult for the teleosemanticist to explain how representational content could arise in the actual world, if she denies that *R* in *C1* has representational content. See Papineau (2001) for another problem with denying that Swampman has representational states.

sustain existential generalization and substitution *salva veritate*. Content satisfying (i)–(iii) then plays the 'right sort' of functional role in Tye's view when it is *poised,* in that it "stands ready and available to make a direct impact on beliefs and/or desires" (2000, 62). Thus, on Tye's view, if a state has PANIC (i.e., *P*oised, *A*bstract, *N*on-conceptual, *I*ntentional *C*ontent), it is phenomenally conscious.

Independently of whether we endorse Tye's view or some other strong representationalist proposal, in order to explain conscious experience, we would still need an account of what it is for a state to have representational content in the first place. That is, strong representationalist theories need to explain how experiences get their content.

Typically, these theories are combined with a reductive, naturalistic theory of content because if such combination is successful, this will have the advantage of allowing for a naturalistically acceptable explanation of conscious experience (see Fish 2010, 77 for details). Even though there are a number of different naturalistic theories of representation available (e.g., Fodor 1990, Whyte 1990, Harman 1987), strong representationalists tend to subscribe to variants of teleosemantics to explain the representational-content part of their view of experience (see, e.g., Dretske 1995, 15; Tye 1995, 153; Lycan 1996, 75). This is because teleosemantics is widely regarded as the most plausible naturalistic account of representation.

However, theories of conscious experience that take representational content to be constitutive of the experience and include teleosemantics (with its commitment to selection history) as their account of content have, given the revised Swampman argument above, the following problematic consequence. According to these theories of consciousness, the creature with which the evolution of conscious experience began couldn't have had any conscious experience even though it was behaviourally and functionally identical to and could have co-existed with creatures that were conscious. The reason for this is the same as the one mentioned above with respect to *C1*. According to the argument above, at the beginning of the evolution of representation, there was an organism in the actual world that was behaviourally identical to its conspecifics that had representational states (its offspring), yet, on the teleosemantic picture, still lacked any representational state itself. Since the theories of consciousness at issue take representational content to be *constitutive* of conscious experience and in addition endorse teleosemantics to explain content, on these theories, the first creature with a conscious state in the actual world, at the beginning

of the evolution of consciousness, was in the same situation as *C1*. It was behaviourally identical to creatures with conscious states (its offspring), yet lacked any conscious state.[9]

The problem with this is that it threatens the causal efficacy of consciousness. For if there are two beings in the actual world that are behaviourally identical but only one of them is conscious, then it seems consciousness does no longer matter causally for the behaviour and survival of these beings. However, conscious experience clearly does affect behaviour and survival. It is, for instance, surely your consciously experiencing the pain that causes you to withdraw your hand from the hot plate. Furthermore, if consciousness didn't have a causal impact on behaviour, it is unclear why it should have evolved in the first place, for it would then not have been able to make any difference to the organism's fitness. Since consciousness did evolve, and does matter causally, the view that conscious experience is causally inert is unacceptable. As a part of strong representationalism, teleosemantics thus leads to the wrong result.

To be clear, there might be representationalist theories of consciousness that do not hold that the representational content is constitutive of conscious experience. There might also be representationalist theories that don't include teleosemantics as an account of content. Or it could turn out that, as a matter of fact, a non-representationalist theory of consciousness is the most tenable view.

While these possibilities remain open, the accounts that I'm focussing on here, namely theories such as Dretske's (1993, 1995) and Tye's (1995, 2000) that do combine teleosemantics with the claim that content is constitutive of the phenomenal character of conscious experience will be faced with the problem of preserving the causal efficacy of consciousness.[10]

Note that while one could have proposed the preceding argument about the efficacy of consciousness already with respect to the fictional Swamp-

---

9. Dretske is in fact ready to bite the bullet and to deny that creatures such as Swampman have conscious states. Tye, however, holds in his first book that Swampman does have conscious states. As it happens, he has changed his mind and now agrees with Dretske. Thanks to a referee of this journal for pointing this out. Below I mention why it is problematic to deny Swampman-like creatures consciousness.

10. The reasoning offered here could be extended to any etiological account of biological function in general. To do so, one only needs to replace *R* and representational content above with a particular trait and one's preferred biological function. The causal-efficacy issue raised here with respect to consciousness, and Dretske's and Tye's theories will then arise with respect to this biological function. The argument will also apply to theories of meaning that tie meaning to indicator function and indicator function to etiology.

man, using *C1* in the argument adds the following twist to it. In response to the original Swampman objection, representationalists such as Dretske and Tye could hold that strong representationalism has the status of a necessary *a posteriori* truth (if true at all). The discovery of Swampmen would then constitute an empirical refutation of strong representationalism. But since there are no Swampmen in this world, Dretske et al. could continue, their account isn't threatened. By replacing Swampman with *C1,* which is an actual creature, this move is now blocked.

IV. *Representationalist responses*

One strategy that strong representationalists such as Dretske or Tye might consider in order to deal with the problem discussed would be to make selection history 'reach' *R* in *C1*. Indeed, there is no reason to suppose that *R*'s selection history can only be construed inter-generationally. For instance, Papineau (1984) and Dretske (1988) speak of learning and conditioning as synchronic, non-genetic selection processes occurring alongside inter-generational natural selection.

However, Papineau's and Dretske's way of specifying intra-generational selection, namely in terms of learning or conditioning is unsatisfactory when it comes to the issue at hand. For, arguably, the representational contents of the most basic sensory-perceptual representations aren't acquired via learning: one doesn't *learn* to represent some red round object as red round object, even though one might learn that the object which one represents as red and round is, say, a tomato. The representational content we are currently interested in, that is, the representation in the first creature to ever have a representational state hence can't be explained by appeal to learning or conditioning. Explaining intra-generational selection in terms of learning or conditioning won't help support the view that *C1* has a state with representational content.

A different way of dealing with *C1* might be found in Tye (1998). He writes that in the Swampman scenario

> there are conditions under which [Swampman] will flourish, and there are conditions under which he will not. If objects in the external environment trigger internal states in Swampman that elicit behaviour inappropriate to those objects—if, say, light rays bend in peculiar ways, thereby causing Swampman to misidentify very badly the shapes and sizes of things—then he isn't going to last long. […] This leads to the thought that Swampman

can have inner states that acquire representational content via the tracking or causal covariation that takes place under conditions of well-functioning. […]

[W]here the representational contents of experiences are concerned, what counts as tracking in normal conditions can vary with the kind of creature or system we are dealing with. Where there is a design, normal conditions are ones in which the creature or system was designed to operate. Where there is no design, normal conditions are, more broadly, ones in which the creature or system happens to be located or settled, if it is functioning well (for a sufficient period of time) in that environment. (1998, 463)

While Tye's proposal looks promising, his idea that states in Swampman (and by extension *C1*) "acquire representational content via the tracking or causal covariation that takes place *under conditions of well-functioning* [my emphasis]" (Ibid) leads to the following problem. Suppose that *R* in *C1* co-varies with *X**, which, as it happens, is an innocuous slithery creature but nevertheless initiates avoidance behaviour in *C1*. Given that it initiates entirely unnecessary avoidance behaviour, *R* will not contribute to *C1*'s well functioning but in fact undermine it by reducing her available energy resources. Since that is so, on Tye's view, *R* will presumably not be about *X** and, assuming we accept Tye's account of consciousness, and are considering *R* in *C1* as the first conscious state, won't be a conscious state, as it will lack content. However, suppose that at some point, *X**s develop a disease that is deadly for *C1*. Avoiding *X**s now does contribute to *C1*'s well functioning. As a result, on Tye's view, *R* will now be about *X**s and (assuming it meets further conditions, see above) be conscious. The problem with this is that *R* in *C1* is with respect to its behavioural and internal effects in the two different scenarios identical,[11] but in Tye's view, in one case *C1* will be conscious while in the other she won't. If that is so, then Tye's account doesn't help avoid the initial problem about preserving the efficacy of consciousness in actual creatures. Furthermore, if the account were right, it seems we could cause *C1* at various times to become conscious and unconscious simply by changing the physical constitution of *X**s—which is hard to accept.

11. To be sure, before *X**s become diseased, *R* in *C1* is detrimental to her well functioning, whereas when *X**s acquire the disease, *R*'s effects will be beneficial. This looks like a significant causal difference between the two scenarios. However, whether *X**s are diseased or not is completely irrelevant for *C1*'s *behaviour*: in both scenarios *R* will lead to the same avoidance behavior when confronted with *X**s, to the same consumption of resources etc. In fact, assuming that *R* always keeps *C1* away from a diseased *X**, *C1*'s course of life will be the same before and after the change in *X**s. If in one case *C1* is conscious and in the other unconscious, then consciousness is no longer causing behavior.

There might be other ways in which strong representationalists such as Dretske and Tye could respond to the original Swampman argument and the revised version that was introduced in this paper. It is, however, not obvious what these responses would be. For the time being, the argument offered doesn't only undermine the fiction response to the Swampman worry, but also poses a significant problem for strong representationalist theories, if they rely on teleosemantics as their account of representational content. On this conjunction of theories (i.e., strong representationalism and teleosemantics), the efficacy of conscious states in creatures in the actual world is undermined. Assuming that consciousness is efficacious, one of the two theories in the combination will need to be modified or abandoned.[12]

References

Adams, Fred and Aizawa, Ken 2010: "Causal Theories of Mental Content". *The Stanford Encyclopedia of Philosophy* (Spring 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/content-causal/>.

Bickhard, Mark 1993: "Representational Content in Humans and Machines". *Journal of Experimental and Theoretical Artificial Intelligence* 5, 285–333.

— 2004: "The Dynamic Emergence of Representation". In: Hugh Clapin (ed.), *Representation in Mind*. Amsterdam: Elsevier, 71–91.

Braddon-Mitchell, David and Jackson, Frank 1997: "The Teleological Theory of Content". *Australasian Journal of Philosophy* 75, 474–489.

Byrne, Alex 2001: "Intentionalism Defended". *The Philosophical Review* 110, 199–240.

Campbell, Donald T. 1974: "Evolutionary Epistemology". In: Paul Schilpp (ed.), *The Philosophy of Karl Popper* Vol. I. Illinois: La Salle, 413–459.

Chalmers, David 2004: "The Representational Character of Experience". In: Brian Leiter (ed.), *The Future for Philosophy.* Oxford: Oxford University Press, 153–182.

Cummins, Robert 1975: "Functional Analysis". *Journal of Philosophy* 72, 741–765.

12. Many thanks to the participants of the TOC group at the University of Alberta, where the main idea of this paper was first discussed in 2011. Thanks also to Martin Lenz (and the participants of the KCL-HU Berlin graduate workshop in 2012), Matteo Mameli, David Papineau, and Janiv Paulsberg for useful discussions. I'm also grateful to Tobias Wilsch, Anna Garrouty, and two anonymous referees of this journal for comments on earlier drafts of the paper.

Cummins, Robert 1996: *Representations, Targets, and Attitudes.* Cambridge, Mass.–London: MIT Press.
— 2002: "Neo-Teleology". In: Andre Ariew, Robert Cummins, and Mark Perlman (eds.), *Functions: New Essays in Philosophy of Psychology and Biology*. Oxford: Oxford University Press, 157–173.
Davidson, Donald 1987: "Knowing One's Own Mind". *Proceedings and Addresses of the American Philosophical Association* 60(3), 441–458.
Dretske, Fred 1981: *Knowledge and the Flow of Information*. Cambridge, Mass.–London: MIT Press.
— 1988: *Explaining Behaviour*. Cambridge, Mass.–London: MIT Press.
— 1993: "Conscious Experience". *Mind* 102, 263–283.
— 1995: *Naturalizing the Mind*. Cambridge. Mass.–London: MIT Press.
Fish, William 2010: *Philosophy of Perception*. New York: Routledge.
Fodor, Jerry 1990: *A Theory of Content and Other Essays*. Cambridge, Mass.–London: MIT Press.
Harman, Gilbert 1987: "(Non-Solipsistic) Conceptual Role Semantics". In: Ernest LePore (ed.), *New Directions in Semantics.* London: Academic Press, 55–83.
Lycan, William 1996: *Consciousness and Experience*. Cambridge, Mass.-London: MIT Press.
Macdonald, Graham 1989: "Biology and Representation". *Mind and Language* 4, 186–200.
Macdonald, Graham and Papineau, David 2006: "Introduction: Prospects and Problems for Teleosemantics". In: Graham Macdonald and David Papineau (eds.), *Teleosemantics: New Philosophical Essays*. Oxford: Oxford University Press, 1–22.
McLaughlin, Brian 2003: "Color, Consciousness and Color Consciousness". In: Aleksandar Jokic and Quentin Smith (eds.), *Consciousness: New Philosophical Perspectives*. Oxford: Oxford University Press, 97–157.
Millikan, Ruth 1984: *Language Thought, and Other Biological Categories.* Cambridge, Mass.–London: MIT Press.
— 1996: "On Swampkinds". *Mind and Language* 11 (1), 103–117.
— 2000: O*n Clear and Confused Ideas: An Essay About Substance Concepts.* Cambridge, Mass.–London: MIT Press.
Neander, Karen 1996: "Swampman Meets Swampcow". *Mind and Language* 11(1), 118–129.
— 2006: "Naturalistic Theories of Reference". In: Michel Devitt and Richard Hanley (eds.), *The Blackwell Guide to the Philosophy of Language.* Malden: Blackwell Publishing, 374–392.
— 2012: "Teleological Theories of Mental Content". *Stanford Encyclopedia of Philosophy*. (Spring 2012 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/entries/content-teleological/.

Papineau, David 1984: “Representation and Explanation”. *Philosophy of Science* 61, 550–572.
— 1987: *Reality and Representation*. Oxford: Basil Blackwell.
— 1993: *Philosophical Naturalism*. Oxford: Blackwell.
— 1998: “Doubtful Intuitions”. *Mind and Language* 11 (1), 130–132.
— 2001: “The Status of Teleosemantics, or How to Stop Worrying about Swampman”. *Australasian Journal of Philosophy* 79 (2), 279–289.
— 2006: “Naturalist Theories of Meaning”. In: Ernest Lepore and Barry C. Smith (eds.), *The Oxford Handbook of Philosophy of Language*. Oxford: Oxford University Press, 175–189.
Shea, Nicholas 2013: “Naturalising Representational Content”. *Philosophy Compass* 8 (5), 496–509.
Stampe, Dennis 1977: “Toward a Causal Theory of Linguistic Representation”. *Midwest Studies in Philosophy* 2 (1), 42–63.
Tye, Michael 1995: *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, Mass.–London: MIT Press.
— 1998: “Inverted Earth, Swampman, and Representationalism”. *Philosophical Perspectives* 12, 459–478.
— 2000: *Consciousness, Color and Content*. Cambridge, Mass.–London: MIT Press.
Whyte, Jamie 1990: “Success Semantics”. *Analysis* 50 (3), 149–157.

Grazer Philosophische Studien
90 (2014), 289–306.

CONTEXTUALISM AND THE PRINCIPLE OF TOLERANCE

Paula SWEENEY
University of Aberdeen

Summary
When we bring together certain plausible and compatible principles guiding the use of vague predicates, the inclination to accept that vague predicates are tolerant is significantly weakened. As the principle of tolerance is a troublesome, paradox inducing principle, a theory giving a satisfactory account of the nature of vague predicates and accounting for the appeal of the sorites paradox without recourse to the principle of tolerance is a worthy addition to the vagueness debate. The theory offered, Contextual Intolerance, draws considerably on Sainsbury's (1996) thesis of the boundarylessness of vague concepts and on the contextualist theories of vagueness offered by Stewart Shapiro (2003, 2006) and Diana Raffman (1994).

1. *Introduction*

The motivating intuition behind the principle of tolerance is that vague predicates are such that, if the predicate applies to an object *n*, then the predicate must also apply to an object that is suitably similar to *n* in the relevant respect. A definition of what it takes for a predicate *'is F'* to be *tolerant* was first given by Crispin Wright:

> Standard Tolerance: "Whereas large enough differences in *F*'s parameter of application sometimes matter to the justice with which it is applied, some small enough difference never thus matters." (Wright 1976, 334)

If we stipulate that a *vague* predicate, e.g. 'is bald', is such that standard tolerance governs competence with use or is at the very least a semantic default, then under certain *sorites-prone* conditions standard tolerance shows itself to be a troublesome, paradox inducing, principle. An example of sorites-prone conditions for 'is bald' is a series of 100,000 men, where #1 has no hairs on his head, #100,000 has 99,999

hairs, and each man in the series has exactly one more hair than his predecessor.

The standard version of the sorites paradox sees a false conclusion reached from apparently true premises via apparently valid reasoning. Consider this example ranging over the series of 100,000 men, where *F* schematises the vague predicate 'is bald'[1],

Base step:	*F1*
Universal Generalisation:	$(\forall x)(Fx \rightarrow Fx')$
Conclusion:	*F100,000*

The base step represents the fact that the first man in the series is bald. The universal generalisation represents our tolerance intuition that, for any object that satifies the predicate 'is bald', an object which is relevantly similar will also satisfy the predicate. Elementary reasoning via universal instantiation and modus ponens leads to the conclusion *F100,000*, yet a man with 99,999 hairs clearly does not satisfy the predicate 'is bald', hence the paradox.

Contextualist solutions to the sorites rely on some form of what Åkerman and Greenough call *weak* tolerance;

> Weak Tolerance: "It is not the case that: there is a context of utterance C and there is an x such that x and x′ are *considered together as a pair by a single subject in C* and 'is *F*' (as used in C) is true of x and 'is *F*' (as used in C) is false of x′, (where x′ is adjacent to x in the sorites series running from *F* to not-*F*)." (Åkerman and Greenough 2010, 276)

Weak tolerance, henceforth simply *tolerance*, has been brought to bear by contextualists in different forms. Diana Raffman represents tolerance with IP*,

> IP*: "for any n, if patch #n is red then patch #(n+1) is red, relative to a pairwise presentational context". (1994, 68)

Delia Graff Fara represents tolerance with Salient Similarity (SS),

> SS: "if two things are saliently similar, then it cannot be that one is in the extension of the predicate, or in its anti-extension, while the other is not." (2000, 57)

---

1. Where *x′* is a complex expression involving the bound variable *x*, equivalent to a description such as 'the successor of *x*'.

And Stewart Shapiro gives the weak contextualist version of Wright's tolerance principle (CT),

> CT: suppose that two objects *a, a'* in the field of *P* differ only marginally in the relevant respect (on which *P* is tolerant). Then if one judges *a* to have *P*, then she cannot competently judge *a'* in any other manner [within the same conversational context]. (2006, 8)

The contextual theorists propose that through context dependent judgements and context shifts they can give a solution to the paradox that has the desirable features of preserving the principle of tolerance and providing an explanation of the seductive appeal of the paradox.[2]

Here it is proposed that the context shifts required to provide this package are otherwise unmotivated and ad hoc. Shapiro's semantic solution to the paradox, arguably the most robust of the contextualist theories, is the focus of the specific criticisms, but the general challenge—that non-standard (i.e. non-Kaplanian) contextual mechanisms must be motivated independently of the fact that they may provide a solution to the sorites paradox—applies to all extant contextualist theories of vagueness.[3] Moving towards a positive contextualist account that requires no ad hoc context shifts, it is proposed that the status of the principle of tolerance as a principle governing the competent use of vague predicates is significantly weakened when it is positioned in a contextualist framework. Vagueness is to be characterised not by tolerance but by elements of permissibility operating within a context.

2. *Contextualism and boundarylessness*

Contextualist theories rely heavily on what are known as 'forced-march' versions of the sorites. In the forced-march sorites, paradox arises when we imagine a judger progressing along a sorites-prone series, like the series of 100,000 men described above, and passing judgements. The judgements are constrained by two incompatible principles that appear to govern the

2. Note that there is a sense in which Graff Fara is the odd one out of the trio: for Shapiro and Raffman, weak tolerance allows us to claim that in no context is there a cut-off, for Graff Fara weak tolerance allows us to claim that the context-dependent cut-off can never be known.

3. The particular challenges I present here are most relevant to Shapiro and Raffman's theories. For a detailed account of the ad hoc features of Graff Fara's contextualist theory see Sweeney and Zardini (2011).

competent use of vague predicates. The first principle states that vague predicates are tolerant—as above—and the second states that a competent judger must respect what Delia Graff Fara (2000) termed the *clear-case* constraint.

> *Clear-case Constraint*: "For each predicate there will be only a limited range of cases which it will be permissible to count as positive instances and there will be a class of things which it will be mandatory to class as positive instances." (Fara 2000, 57)

Consider the series of 100,000 men, proceeding from a clearly bald man (#1) to a clearly not bald man (#100,000), with each man in the series differing only marginally from his predecessor in terms of the number and arrangement of hairs. Assuming tolerance is in play, it is easy to see how the forced-march paradox will arise. In order to be competent one must respect the clear-case constraint, judging #1 to be bald and #100,000 to be not bald, and yet we are, we shall assume, guided by the principle of tolerance which states that at each stage, if x is judged to be bald then x', if judged at all, must be judged likewise. Given these restraints, the forced-march game is one that we cannot win: if we respect tolerance we violate the clear-case constraint, if we respect the clear-case constraint we violate tolerance.

According to Shapiro, we must "jump" at some point when going through the series, from *x is bald* to *x is not bald* judgements, in order to avoid the incompetent claim that #100,000 is bald—but the question is, how can we do this without violating tolerance?[4]

First, we can narrow down the area where such a jump would be permissible, notwithstanding tolerance for the present. The series contains *determinate* cases for which the facts about an object x (such as the number and arrangement of hairs) determine whether the conditions of application of the predicate have been met. In such cases it is determinate that a sentence, S, which asserts that x has the property F is either true or false. However, somewhere in the series there is a borderline case: a case for which it is neither determinate that S is true nor determinate that S is false.[5] For any such case, competent assertions of S are governed by the principle of 'open-texture':

---

4. Notice that the claim that judgers will jump from 'bald' to 'not bald' judgements is a simplification; a reluctance to judge or a 'not sure' assertion is also classed as a jump.

5. Shapiro states that such a borderline case is one in which the *non-linguistic* facts have not determined the truth-conditions for '*Fa*'. However, as Greenough (2005, 170) notes, this

> *Open-texture*: In borderline cases, a speaker is free to assert '*Fx*' or '¬*Fx*', without offending against the meanings of the terms, or against any other rules of language use. (2006, 10)

Shapiro also employs the following response-dependence principle which determines the *extension* of the vague predicate:

> Judgement-dependence: Vague predicates are judgement-dependent (in the borderline area) such that an item lies in a given category iff the relevant subjects would judge it to lie in that category. (2006, 41)

Shapiro incorporates the open-texture thesis into a conversational framework that allows for an assertion regarding an open case to be, not only assertable, but *true* in a context, where truth in *a* context is distinct from truth in *every* context (determinate truth). This notion of truth in a context opens up a possibility. If Shapiro can show that a jump brings about a change in conversational context such that an assertion of *Fx* and an assertion of ¬*Fx*′ occur in *different* contexts, the jump can occur without violation of tolerance. Requiring some mechanism to regulate this proposed feature of language, Shapiro turns to Lewis's theory of conversational scorekeeping (Lewis 1979).

Lewis claims that in any conversation a 'scorecard' will record such things as shared assumptions, details of the environmental context, salient objects or persons, comparison classes, and other information that is required to determine which conversational moves are permissible. This record is necessary for determining the assertability conditions of a given statement within the conversation and evolves continually with each assertion or change in the environment.

Shapiro claims that this is just the framework that his context-change theory requires. If he can persuade us that a jump from the assertion of the sentence *That (x) is bald* to the assertion of *That (x′) is not bald* brings about a change in conversational context and in particular that the addition of *That (x′) is not bald* to the conversational record *removes* the earlier *That (x) is bald*, weak tolerance can be protected and the paradox avoided.[6] The analogy with Lewis's use of the mechanism is given:

---

is not what Shapiro *should* say as, for Shapiro, a vague predicate is one where the *conditions of application* are incomplete. That is, it is the linguistic facts and not the non-linguistic facts which fail to determine the truth conditions for '*Fa*' in a borderline case.

6. See Gross (2009) for a discussion of the tension between tolerance and open-texture. Gross also clarifies that although Shapiro does not claim that tolerance is a part of a vague term's meaning it is consistent with his view that to maintain tolerance is a semantic default. (2009, 262, fn 1)

> The event described here is quite similar to the outcome in one of Lewis's scenarios […] In that story, the participants in a conversation first agree to a 'low' standard when they accept 'France is hexagonal'. Later, when they demur from 'Italy is boot shaped', the standard is raised, and so 'France is hexagonal' is implicitly removed from the record. Similarly, when the present conversationalists explicitly declare that #975 is not bald, they implicitly retract the statement that #974 is bald. In short, the conversational score is the device that enforces tolerance. (2003, 52)

According to Shapiro, not only is *number 974 is bald* removed but also a 'backward spread' (Raffman, 1994) occurs removing an indeterminate number of earlier assertions from the conversational record.[7] In any case, the removal of (at least) the earlier assertion, along with Shapiro's assumption that tolerance does not hold *across* contexts, allows that x' can be judged to fall outwith the extension of the predicate and x judged to fall within the extension, without violation of tolerance and without compromising competence. The change of context accompanying the jump permits the judger to pass legitimately from assertions of the sentence *That is bald* to assertions of the sentence *That is not bald* within the borderline area.

Shapiro's solution to the forced-march sorites depends heavily on its being plausible that the context shifts in the way that he describes. However, while Lewis is explicit about which parameters shift, what causes them to shift and, most importantly, how participants in the conversation are aware of such shifts, Shapiro is silent. Shapiro's theory is at risk of appearing ad hoc.

Shapiro must motivate the view that an assertion of the sentence *That is bald*, of an object in the borderline area, and an assertion of the sentence *That is not bald* of an adjacent object creates an unacceptable tension on the scorecard. At first glance, such a tension does not seem obvious: there is nothing *contradictory* about *That (x) is bald* and *That (x′) is not bald*. But perhaps the tension that Shapiro requires can be created.

What follows is not exegesis of Shapiro's theory but an attempt to search for the kinematics of scorekeeping in the context of a forced-march scenario. There are (at least) two possible approaches, both supported by some textual evidence, which Shapiro may consider to be obvious candidates.

Perhaps Shapiro assumes a tension is created on the scorecard between a presupposition and an assertion. This is a plausible option as it would

7. Shapiro borrows Raffman's psychological account of the phenomenon of backward spread as the basis of his semantic version. See Raffman (1994, 178–80).

account for Shapiro's assumption that the existing Lewisian framework is already equipped for his theory, given that the Lewisian framework is already equipped to accommodate presuppositions.

It may be that, given tolerance, when the objects in the series are connected by a similarity chain, a competent assertion of 'Number 939 in the series is not bald' *requires* the presupposition that number 938 in the series is not bald, such that the assertion of 'Number 939 in the series is not bald' implicitly adds *Number 938 in the series is not bald* to the conversational record. This would create a tension between the presupposed *Number 938 in the series is not bald* and the asserted *Number 939 in the series is bald*. Shapiro alludes to this (2003, 52):

> In declaring man 975 to be not bald, they implicitly deny that 974 is bald, and so 'Man 974 is bald' is removed from the conversational record.

To set up the first problem for this approach suppose that there are two incompatible (by Shapiro's lights) *assertions* (as opposed to an incompatible assertion and presupposition) on the scorecard; *Number 938 in the series is bald* and *Number 939 in the series is not bald*. According to Shapiro, the latter assertion causes a context shift that wipes the earlier assertion (plus a few more) off the scorecard, hereby preserving tolerance. But why should we suppose that it is the *latter* assertion that dominates the scorecard and not the *earlier* one? The fact that there are 'incompatible' assertions on the scorecard gives no notion of *priority*.[8] Neither the scorecard nor the principle of tolerance tells us whether we are in a context in which we are using the predicate $bald_c$ or a context in which we are using the predicate $bald_{c'}$, all we know is that we cannot be employing both predicates in the same context. So, while Shapiro assumes that the earlier assertions are wiped off, the scorecard would be just as efficient in protecting tolerance were it to override the latter assertion.[9]

This problem of direction may seem insignificant but it arises with force with respect to a conflict between an assertion and a *presupposition*. The scorecard contains the asserted *Number 938 is bald*. When the jump occurs and 'Number 939 is not bald' is asserted we are to suppose that the presupposition *Number 938 is not bald* is registered on the scorecard. The

8. As an aside, the same criticism could be made of Lewis—why suppose that higher standards dominate lower standards? Why can lower standards not just remain fixed?

9. It is true that the judger could not progress right through the series without switching predicates, otherwise he would judge at least one clear case incompetently, but that in itself does not justify backwards spread. It is the desire to preserve tolerance that motivates backwards spread.

result is that we have a clash between the asserted *Number 938 is bald* and the presupposed *Number 938 is not bald*. Nothing about the scorecard or the principle of tolerance dictates which of these must change. Furthermore, the items in 'conflict' are an *assertion* and a *presupposition*. Shapiro's view requires that the presupposition would trump the assertion on the scorecard, yet it seems that we have no principled reason for favouring a conflicting presupposition over an assertion.[10] So, even if Shapiro could explain how and why later assertions can override earlier assertions, we would need a further argument from Shapiro to support the claim that later *presuppositions* can override earlier assertions.

The second problem with the presupposition response is that, unless Shapiro offers a distinct theory of presupposition, the relationship between the assertion and the proposed presupposition is not the correct one. According to Stalnaker (whose theory Lewis acknowledges as the basis for his own scorekeeping account of presupposition), a sentence *P* presupposes *Q* if *Q* is required by *both* the assertion of *P* and the assertion of ¬*P*.[11] Consider the following example,

P: The Queen of England is rich.

Both the assertion that *P* and the denial of *P* require the presupposition, *Q*, for assertibility or truth.

Q: There is one and only one current Queen of England.

If we look back now at the sorites case we can see that this relationship does *not* hold between the asserted

R: Number 975 is not bald

and the posited presupposition

S: Number 974 is not bald.

That is, it is not the case that both the assertion that *R* and the assertion that ¬*R* require the presupposition *S*. What Shapiro needs, rather,

10. A desire to honour both the clear-case constraints and the principle of tolerance is not a principled reason.

11. See Stalnaker (1999, 54f.)

is that the assertion *R* requires the presupposition *S* and the assertion that ¬*R* requires the presupposition ¬*S*. Therefore 'Number 974 is not bald' does not have the correct relation to the asserted 'Number 975 is not bald' for it to be implicitly placed on the conversational record as a presupposition. The relation that Shapiro needs between *R* and *S* is that of mutual entailment, not presupposition. But entailment is not available to him.

In searching for the kinematics of scorekeeping in order to bolster the plausibility of Shapiro's scorekeeping mechanism, it seems that we can discount the option of creating a tension between an assertion and a presupposition.

Perhaps the scorekeeping account is to be supported by our *dispositions* to judge: the earlier assertion 'Number 974 is bald' should be removed from the conversational record when 'Number 975 is not bald' is asserted because at that point in the conversation the judger would be *disposed* to deny that number 974 is bald, were he asked. Recalling the phenomenon of backward spread, not only would the judger be disposed to deny that number 974 is bald but he would also be disposed to deny that some of the proceeding members of the series are bald, so the corresponding earlier assertions (a vague amount of them) would be removed from the conversational record. Shapiro seems to support this dispositional account (2003, 52):[12]

> We assume that man #975 is the first 'jump' [...] Suppose that we explicitly ask them about #974 again, after reminding them that they just called that man 'bald', and that they can barely distinguish #974 from #975 (if at all). […] I'd speculate that they would explicitly retract that judgement, saying that #974 is not bald […] Just as 'Man 974 is bald' comes off the record, so does 'Man 973 is bald'; ditto for a few more of their recent pronouncements.

First off we can note that, by going this way, Shapiro avoids the problem of direction as, in this case, the jump itself brings about a change in dispositions regarding the *earlier* judgements.

However, there is something left unexplained here. Shapiro has stated that, for any competent judger, an assertion regarding an object in the series

12. Notice, Shapiro is not claiming that explicit retraction is required for context change. The earlier assertion is removed when a jump occurs even if no retraction is made. Shapiro is simply motivating his theory by saying that the assertions should be removed because, if asked, the subject would retract his earlier assertion.

is accompanied by a disposition to co-classify the objects on either side.[13] At the same time, Shapiro claims that to retain competence a judger must jump at some point throughout the series; in order to do so it is clear that the judger must momentarily override his disposition to co-classify (the tolerance disposition). Yet tolerance must be immediately reinstated via the judger's dispositions as soon as the jump has occurred. That is, it is not simply that the tolerance disposition reasserts itself with regard to *future* judgements but also with regard to recently asserted judgements. What is it that makes the tolerance disposition reassert itself, adjusting recently asserted judgements, the very instant a jump is made?[14] The apparatus is left unexplained.

That worry aside, how does the claim that the conversational record reflects our *dispositions* fit with the semantic principles of tolerance and judgement-dependence? A dispositional version of tolerance would state that a (competent) judger who asserts 'Number 975 is not bald' is disposed to deny that number 974 and number 976 are bald. But note that the linguistic version of backward spread that Shapiro employs goes beyond dispositional-tolerance, as stated. What Shapiro requires is some form of extended dispositional-tolerance: a (competent) judger who asserts 'Number 975 is not bald' would be disposed to assert '*x* is not bald' of a *few* of the objects each side of number 975. This extended dispositional principle will show itself in a dispositional version of Shapiro's judgement-dependence principle. Shapiro's judgement-dependence principle states that an object satisfies a vague predicate iff it is (competently) judged to have *F*. But if Shapiro is committed to the dispositional version of this principle, i.e., that an object satisfies a vague predicate iff a competent judger is *disposed* to judge it to satisfy the vague predicate, then a serious problem arises.[15] The guiding principle behind Shapiro's theory is that

13. This does not entail that the judger has a disposition to judge all of the objects to be the same colour; there is a gap between the disposition $\forall x(J{:}Fx \rightarrow \forall J{:}Fx')$ and the disposition $J{:}Fa \rightarrow \forall x(J{:}Fx)$.

14. Perhaps Shapiro can borrow Raffman's psychological, categoriser/discriminator distinction (see section 3 below and Raffman, 1994, 47), but then tolerance stops being a semantic principle, as Shapiro requires it to be, and instead becomes a psychological principle.

15. Shapiro might respond by claiming that he is committed to a view under which the scoreboard registers speaker's dispositions only and not the dispositional view of judgement that I propose on his behalf. However, as he is committed to vague predicates being dependent, he would still be required to give some account of the relationship between dispositions and judgements. Note also that Shapiro commits himself to a form of judgement-dependence under which consensus of judgement is required, and this certainly seems to be in tension with the dispositional account.

tolerance can be preserved if a jump marks not a boundary but a change in conversational context. But now it seems that when a jump occurs, the dispositions of the judger register on the scorecard, causing a boundary to appear. Consider a judger who, having asserted 'Number 974 is bald', jumps and asserts 'Number 975 is not bald'. According to Shapiro, at least one of the earlier assertions is removed from the conversational record, in accordance with a change in the judger's dispositions. Shapiro supposes that his weak tolerance principle will keep him out of trouble here; if (at least) 'Number 974 is bald' is removed, number 974 is *unjudged* and the principle of tolerance holds. But if in order to remove the asserted 'Number 974 is bald' the scorecard must register the judger's new disposition that number 974 is *not* bald, then 'Number 974 is not bald' registers on the scorecard and tolerance is violated: the scorecard now registers that number 973 is bald and that number 974 is not bald. The extended versions of the principles that accommodate the phenomenon of backwards spread do not help as they simply place the cut-off further down the series.[16]

Shapiro states that the 'conversational score is the device that enforces tolerance' but it is not clear what that amounts to. If Shapiro means that an attempted violation of the principle of tolerance causes the context to shift, we are still lacking any account of the kinematics of scorekeeping that supports this stipulation. In the standard Lewisian cases we explain apparent incompatibility *away* by saying that the value taken by some parameter can legitimately vary across the two contexts; we have one variable standard governing permissible assertions. In the Lewisian case a variation in standard marks a change in context; furthermore, no single context could be both a high and low standard context. In the sorites case, by contrast, it is not that some parameter has shifted during the conversation but rather that there are two incompatible principles governing permissible assertions in each context; the clear-case constraint and the principle of tolerance. The stipulation of the context shift that is required to preserve tolerance in a context is entirely ad hoc. It is difficult to see how to make the conversational mechanism work, except by stipulation.

16. This same criticism can be put to Raffman.

3. *Contextual intolerance*

The contextualists stipulate context shifts to preserve tolerance while accommodating the clear-case constraint. But why think that the principle of tolerance is a restriction on competence with a context dependent vague predicate in any case?

If tolerance is in play during a forced-march, judgers have two options: they can either respect tolerance and judge a clearly not bald man to be bald, or they can violate tolerance and, for some adjacent members, judge the first but not the second to lie in the extension of the predicate, while judging the clear cases correctly.

Regarding the first option, as Shapiro rightly states, "That way lies madness (or at least incompetence)" (2003, 51). A judger who continues right through the series and judges a man with a full head of hair to be bald has displayed incompetence with the term 'bald'. And notice that offering up the principle of tolerance as justification for his judgement in no way eradicates or lessens this incompetence.

Regarding the second option, violating tolerance, what lies that way? Certainly not madness nor, it will now be pressed, even incompetence.[17] Consider the forced march scenario again and imagine a judger who, upon jumping, is not inclined to retract his earlier judgements, instead insisting that #974 is the last bald man and #975 the first non-bald man *within that context*. Would we judge such a judger to be *incompetent*? The judger takes the inclusion of the polar cases to require him to judge the last bald man and first non-bald man to be somewhere in the borderline area. To reason in such a manner within any sorites context would be to reason competently, so why are we to assume that tolerance is the semantic default?

It will prove useful when considering vague predicates to think of them not in negative terms nor in terms of restrictions that may accompany competent usage but in terms of permissibility; what are we permitted to do when applying vague terms? It is in this spirit that Mark Sainsbury (1991) offered the slogan "vagueness offers freedom".

Sainsbury proposes that vague predicates correspond to vague concepts, which are themselves concepts without boundaries. In contrast to a concept being tolerant, a concept being boundaryless does not preclude boundaries being drawn. While boundaryless concepts *need* not draw boundaries

17. Notice that here the judger *can* justify his judgements by offering up the clear-case constraint.

it does not follow that they *must* not draw boundaries. To think that the boundarylessness of vague concepts is restrictive in this way is to make a faulty inference.

> A boundaryless concept is one which, for closely similar pairs, never makes it mandatory to apply the concept to one member of the pair, and withhold it from the other; hence, the argument runs, a boundaryless concept is one which, for closely similar pairs, makes it mandatory never to apply the concept to one member of the pair, and withhold it from the other. The inference depends upon the move from something being not mandatory to its being forbidden; a move legitimate within the totalitarianism of boundary-concepts, but not within the liberality of boundarylessness. (Sainsbury 1990, 260)[18]

While vague concepts *need* not be bound, they can be. Furthermore, a competent judge is aware of this, as grasping a vague concept involves the realisation that "vagueness offers freedom […] one may behave consistently with the nature of the concept in drawing a line between adjacent pairs". (Sainsbury 1990, 259f.) It is this and *not* tolerance which is constitutive of vague predicates.[19]

Sainsbury raises two points of interest. First, he denies that tolerance governs competence with vague predicates: it does *not* follow from the boundarylessness of vague concepts that, for every x, if x has been (competently) judged to satisfy a vague concept, then, in order to retain competence, it is *mandatory* that its successor, x', if judged at all be judged likewise. What follows from the boundarylessness of a vague concept is that for every x, if x has been judged to satisfy a vague concept, then it is *permissible* that its successor, x', if judged at all be judged likewise. This brings us to the second point, Sainsbury's dictum that vagueness offers freedom.

Just how much freedom does vagueness offer? Clearly vagueness offers more freedom than many theorists have realised, the regularly adopted tolerance principle being overly restrictive. On the other hand, the competent use of vague predicates is not entirely unrestricted: there are clear-case and other penumbral constraints (Fine 1975).[20]

---

18. It is interesting that Delia Graff Fara (2000) uses tolerance and boundarylessness interchangeably in her theory of vagueness, yet tolerance does not follow from boundarylessness.

19. Note that, according to Sainsbury, the fact that we are permitted to use vague predicates and draw sharp boundaries is not a feature of pragmatics: permissibility is a feature of the concept, not a pragmatic feature.

20. A *penumbral connection* is a logical or conceptual relation holding among sentences that contain predicates with borderline cases. For example, something which is '*F*-er' than something

Suppose that the freedom emerges from two features. First, as Shapiro's open-texture principle above states, in borderline cases a speaker is free to assert '*Fx*' or '¬*Fx*' without relinquishing competence.[21] Second, as Sainsbury recognises, a judger is permitted to draw (non-legislative) boundaries between adjacent pairs and retain competence.

Accordingly, we can replace the principle of tolerance with the following dual principles which we can collectively refer to as principles of Permissible (In)tolerance:

> *Permissible Tolerance*: For any non-clear (i.e. borderline) case *x* in an ordered series, if *x* is competently judged to satisfy *F* then *x*′ may be judged likewise, if judged at all.
>
> *Permissible Intolerance:* For any non-clear (i.e. borderline) case *x* in an ordered series, if *x* is competently judged to satisfy *F* then *x*′ may be judged to not satisfy *F*, if judged at all.[22]

Rather than stipulating ad hoc context shifts, we assume that the standard Kaplanian context <*agent, location, world, time*> also contains a judge parameter. Variation in any parameter of the context marks a context shift. Most relevant for our purposes will be variation in the objects of the series (including their order of presentation) or in the judger of the context (either with a different judger or with the same judger on a different occasion).

The replacement of the tolerance principle with a permissibility principle allows us to give a minimal contextualist theory of vagueness without

---

which satisfies 'is *F*' must itself satisfy 'is *F*'; also, the same object cannot both satisfy and not satisfy 'is *F*' (within the same context). For more on penumbral connections see Fine (1975).

21. It is perhaps worth raising a general worry here regarding the distinction between *mandatory* cases and *judgement-dependent* or *permissible* cases. It is claimed that vague predicates lack boundaries in that there is an area of permissibility. It does not follow from this claim that there is a switch of the direction of dependence. Shapiro claims that in the clear cases we act as 'detectivists', the clear cases determining our judgements, yet in the non-clear cases we act as 'projectivists', our judgements determining whether or not the predicate applies. It seems feasible that we could go 'projectivist' all the way, defining the clear cases as those which all competent judgers would judge to be clear cases as follows: *x* is a clear case of *F* iff for any subject meeting conditions C, if the subject were to judge *x*, he would judge *Fx*. Thinking of clear cases in this way seems to indicate that there could be a deeper connection between judgement-dependence and vague predicates, but this is too big a digression to be taken up in this paper.

22. Here and throughout we shall stipulate that judgements must reflect the ordering of the series thereby preserving monotonicity.

resorting to ad hoc context shifts. Call such a minimal theory Contextual Intolerance.[23] The principles that form the basis of Contextual Intolerance are the assertion governing open-texture,

> *Open-texture*: In borderline cases a subject is free to assert 'Pa' and free to assert 'not-Pa', without offending against the meanings of the terms, or any other rules of language use,

a contextual version of the extension determining principle of judgement-dependence,

> *Judgement-dependence (in context)*: An item in the borderline area lies in a given category, relative to context *c*, iff the relevant competent subject would judge it to lie in that category in context *c*,

and our principles of Permissible (In)tolerance:

> *Permissible Tolerance*: For any non-clear (i.e. borderline) case *x* in an ordered series, if *x* is competently judged to satisfy *F* then *x′* may be judged likewise, if judged at all.[24]

> *Permissible Intolerance:* For any non-clear (i.e. borderline) case *x* in an ordered series, if *x* is competently judged to satisfy *F*, then *x′* may be judged to not satisfy *F*, if judged at all.

Given Permissible (In)tolerance, and assuming the judgement dependence of vague predicates, the Universal Generalisation is false. It is simply not the case that if any object in a sorites series satifies a vague predicate, an object which is relevantly similar *will* also be judged to satisfy the predicate; rather we say that if any object in a sorites series satifies a vague predicate, an object which is relevantly similar *may* also be judged to satisfy the predicate. The context dependent nature of the judgement-dependence principle

23. The proposed theory is neutral on the question of whether standard contextualism, non-indexical contextualism, truth-relativism or content-relativism gives the correct semantics for language involving vague predicates. However, it is worth noting that any indexical theory, a theory under which content and not just truth value varies with context, will fall foul of Stanley's (2003) verb ellipses sorites.

24. Note that the principle of permissible tolerance must range over pairs of borderline objects in order to make it compatible with the clear-case constraint.

determines that truth of such judgements is context dependent; the relevant propositions are true relative to a context of judgement. That is, '#938 is bald' may be true relative to judgement context *c* but false relative to judgement context *c′*. On this view, context shifts are employed to *accommodate* the variable extensions of vague predicates, not to explain them away.[25]

Have we 'solved' the sorites? It is certainly the case that tolerance should be replaced by Permissible (In)tolerance. It is also the case that Permissible (In)tolerance does not entail the Universal Generalisation, unlike standard tolerance. But it is another question whether or not this satisfies us with regard to the paradox. The paradox has great appeal that is difficult to explain away. Below are considerations that take us at least some of the way towards this goal.[26]

The appeal of the paradox is partly explained by our propensity to confuse the lack of acontextual boundaries with a lack of contextual boundaries. Contextual Intolerance points out that vague predicates permit context-dependent cut-offs.[27] This differs from the claim that there is a cut-off such that *that very cut-off* holds in every context, which the Contextual Intolerance theorist denies. That is, we can accept the following contextualised version of the sharp boundaries claim, $(\exists c)(\exists x)$ $(Fx\text{c} \,\&\, \neg Fx'_c)$, and reject the acontextual version, $(\exists x)(\forall_c)(Fx_c \,\&\, \neg Fx'_c)$.

It is also very plausible that the false mandatory tolerance principle,

> *Mandatory Tolerance*: $\forall x\ J{:}Fx \rightarrow$ (if judged at all) $\boldsymbol{MJ}{:}Fx'$
> (for every *x*, if *x* has been competently judged to satisfy a vague predicate, then, in order to retain competence, it is *mandatory* that its successor, *x′*, if judged at all be judged likewise),

has been confused with the nearby true permissible tolerance principle,

> *Permissible Tolerance*: $\forall x\ J{:}Fx \rightarrow \boldsymbol{PJ}{:}Fx'$
> (for every *x*, if *x* has been judged to satisfy a vague predicate, then *x′* *may* be judged likewise).

---

25. We are not always limited to reasoning pair by pair. Cannot a judge who judges equally two similar objects that are borderline cases of "F" reason correctly by universal generalisation and fall into the soritical conclusion? No. For the judge's dispositions have determined a cut off within the context, making reasoning via universal generalisation faulty reasoning.

26. Note that the diagnosis of the appeal of the sorites offered here (although not the solution) is perfectly analogous to that of the supervaluationist. (See Fine, 1975.)

27. In this respect Contextual Intolerance is similar to Graff Fara's (2000) view.

These principles are in close cognitive proximity, but only one of them leads us to paradox.

4. *Conclusion*

Contextualist theories of vagueness have attempted to provide us with a package in which context dependent judgements and context shifts allow for preservation of the principle of tolerance and of the clear-case constraint. Unfortunately the context shifts that must be stipulated in order to preserve the principle of tolerance are entirely ad hoc and otherwise unmotivated. However, we can have something very close to the desirable package once we realise that it is a mistake to think that mastery of vague predicates is governed by a principle of tolerance. Rather mastery of vague predicates is governed by the (nearby) permissible (in)tolerance principle, a principle which does not lend itself to paradox.[28]

References

Åkerman, Jonas & Greenough, Patrick 2010: "Hold the Context Fixed, Vagueness Still Remains". In: Richard Dietz & Sebastiano Moruzzi (eds.), *Cuts and Clouds: Vaguenesss, Its Nature and Its Logic.* Oxford: Oxford University Press, 275–288.

Fara, Delia Graff 2000: "Shifting Sands: An Interest-Relative Theory of Vagueness". *Philosophical Topics* 28, 45–81. (Originally published under the name "Delia Graff").

Fine, Kit 1975: "Vagueness, Truth and Logic". *Synthese* 30, 265–300.

Greenough, Patrick 2005: "Contextualism about Vagueness and Higher-Order Vagueness". *The Aristotelian Society* 79, 167–190.

Gross, Steven 2009: "Review of Shapiro's 'Vagueness in Context'". *Philosophical Review* 118(2), 261–266.

Lewis, David 1979: "Scorekeeping in a Language Game". *Journal of Philosophical Logic* 8, 339–359.

Raffman, Diana 1994: "Vagueness without Paradox". *Philosophical Review* 103, 41–74.

28. Thanks to an anonymous referee for this journal, whose comments led to vast improvements of the paper in places and aspects too numerous to mention individually.

Sainsbury, Richard Mark 1991: "Concepts without Boundaries". Reprinted in: Rosanna Keefe & Peter Smith (eds.) 1996, *Vagueness: A Reader*. Cambridge, Mass.: MIT Press, 251–264.

Shapiro, Stewart 2003: "Vagueness and Conversation". In J.C. Beall & Michael Glanzberg (eds.), *Liars and Heaps*. Oxford: Oxford University Press, 39–72.

— 2006: *Vagueness in Context*. Oxford: Oxford University Press.

Stalnaker, Robert 1999: "Assertion". In: Robert Stalnaker (ed.), *Context and Content.* Oxford: Oxford University Press, 47–62.

Stanley, Jason 2003: "Context, Interest-Relativity, and the Sorites". *Analysis* 63, 269–280.

Sweeney, Paula & Zardini, Elia 2011: "Vagueness and Practical Interest". In: Paul Egre & Nathan Klinedinst (eds.), *Vagueness and Language Use*. Hampshire: Palgrave, 249–282.

Wright, Crispin 1976: "Language Mastery and the Sorites Paradox". In: Gareth Evans & John McDowell (eds.), *Truth and Meaning: Essays in Semantics*. Oxford: Oxford University Press, 223–247.

Grazer Philosophische Studien
90 (2014), 307–316.

WAS HEISST „SICH VORSTELLEN, EINE ANDERE PERSON ZU SEIN“?

Tammo LOSSAU
Universität Göttingen

Gewinner des ersten Preises des Essay-Wettbewerbs für Studierende 2013
gesponsert durch die Gesellschaft für Analytische Philosophie (GAP)
in Kooperation mit den Grazer Philosophischen Studien*

Summary

Talking about “being another person”, many different things may be meant. I make use of Wollheim’s distinction between three different modes of imagination and invoke four different kinds of possible content of what may be imagined. In effect, I aim at a hopefully complete overview of the possible imaginative projects of “imagining being another person”. I try to keep an eye on the role of numerical identity in each case.

Wenn man sie auffordert: „Stell dir vor, du wärst eine andere Person!“, dann behaupten die meisten Leute, sie könnten sich das *zumindest bis zu einem gewissen Grad* vorstellen. Was genau sie sich dabei allerdings vorstellen, ist sehr unterschiedlich. Ich möchte die verschiedenen denkbaren Reaktionen auf diese Aufforderung systematisch durchgehen und unterscheide dabei erstens drei *Modi* des Vorstellens und zweitens vier *inhaltliche Verständnisse*, was mit „eine andere Person sein“ gemeint sein könnte. Es ergeben sich so zwölf prinzipiell denkbare Typen *imaginativer Projekte*,[1] die jedoch nicht alle sinnvoll möglich sind.

---

* Die Preisfrage des Wettbewerbs 2013 war: „Kann ich mir vorstellen, eine andere Person zu sein?“ Unter den 36 Einreichungen, die den Regeln des Wettbewerbs entsprachen, wählte die Jury drei Beiträge aus, die den jeweils ersten, zweiten und dritten Platz belegten. Die Autorinnen und Autoren der Gewinnerbeiträge erhielten die Erlaubnis, ihre Essays für die endgültige Publikation geringfügig zu überarbeiten.

1. Der Begriff ist von Bernard Williams (vgl. Williams 1973, erstmals S. 30). Ich möchte den Begriff hier so verwenden, dass ein imaginatives Projekt durch seinen Modus sowie seinen beabsichtigten Inhalt (also einschließlich seines Objektes) individuiert wird.

Zunächst unterscheide ich nach Richard Wollheim zwischen *ich-zentriertem*, *azentriertem* und *peripherem Vorstellen*. Die Verständnisse von „eine andere Person sein" unterscheide ich zum Teil nach Peter Goldie folgendermaßen: Angenommen, Anna soll sich vorstellen, Ben zu sein, dann könnte Anna versuchen

(1) sich vorzustellen, sie selbst wäre an Bens Stelle, und sich fragen, wie sie sich fühlen und verhalten würde (*In-deinen-Schuhen-Stecken*), oder
(2) sich empathisch in Ben *einzufühlen* und sich fragen, wie es für Ben ist, er zu sein, oder
(3) die *epistemische Möglichkeit* zu erwägen, dass Anna und Ben *tatsächlich* identisch sind (*Identität in gerader Linie*), oder
(4) sich *kontrafaktisch* eine Person als *Vereinigung* der beiden in Wirklichkeit nicht-identischen Personen Anna und Ben vorzustellen (die der *metaphysischen Möglichkeit* derer Identität entspricht).

Ich möchte im Folgenden die Unterscheidung zwischen den Modi des Vorstellens genauer erläutern (I). Anschließend werde ich die vier angedeuteten Verständnisse durchgehen und erläutern, was dabei genau geschehen soll (II–V). Am Schluss (VI) fasse ich noch kurz zusammen. Ich versuche so zu klären, in welchem Sinne wir uns vorstellen können, eine andere Person zu sein, und in welchem nicht.

I.

Zunächst also gehe ich auf die Modi des Vorstellens ein. Dazu verwende ich Begriffe von Wollheim (vgl. Wollheim 1984, 71–76), die ich aber anhand eines Beispiels von Goldie einführen möchte:

> Let us say that whilst sitting at my desk I imagine the following: I am swimming in what I know to be waters which contain jellyfish and other dangerous creatures. I swim into something slippery (is it just seaweed?) which grips my ankles and impedes my stroke. I strike out for the shore, sensing the grip of the slippery thing getting firmer. I gulp a huge mouthful of salt water. I realize that I am beginning to lose my strength and to feel panicky … (Goldie 2000, 195)

Hier stelle ich mir nur vor, ich selbst sei in einer anderen Situation. Dennoch habe ich unterschiedliche Möglichkeiten, mir dies vorzustellen. Erstens – was bei dieser Schilderung des Szenarios nahe liegt – kann ich mir vorstellen die beschriebene Situation als Akteur zu erleben. Dazu gehört, dass ich die Hilflosigkeit und Panik am Schreibtisch nacherlebe und mir selbst die Gedanken

zu eigen mache, die mir (dem Schwimmer) durch den Kopf gehen. Diese Art des Vorstellens ist *zentriert*, und zwar bezogen auf mich, den Schwimmer (im Folgenden *ich-zentriert*). Zweitens kann ich mir aber die Situation auch von außerhalb vorstellen. Hier sind zwei Fälle denkbar: Einmal kann ich mir die Situation *peripher* vorstellen, und zwar aus der Perspektive einer Außenstehenden (z.B. einer vorbeischwimmenden Taucherin), die beobachtet, wie ich mich im Seegras (oder in Quallen) verfange, oder aber ich kann sie mir *azentriert*, d.h aus einer nicht-personalen Perspektive vorstellen – gewissermaßen als allwissender Erzähler.

Peripheres Vorstellen ist systematisch gesehen ein Spezialfall von zentriertem Vorstellen. Auch hier stelle ich mir vor, jemand zu sein und etwas wie diese Person zu erleben – nur eben nicht diejenige Person, um die es in meinem imaginativen Projekt eigentlich geht. Allerdings hat peripheres Vorstellen wichtige Eigenschaften mit azentriertem Vorstellen gemeinsam: In beiden Fällen kann ich einerseits Dinge wahrnehmen (oder wissen), die dem eigentlichen Objekt der Vorstellung entgehen. So kann ich mir aus diesen Perspektiven vorstellen, dass sich meinem Schwimmer-Ich von hinten eine Qualle nähert, ohne dass es dies bemerkt. Das Miterleben der Emotionen, Wünsche und Wahrnehmungen des Vorstellungsobjektes gehört andererseits in diesen beiden Perspektiven nicht zur Vorstellung. Zwar stelle ich mir vor, zu sehen, wie ich, der Schwimmer, in Panik gerate, aber diese Panik muss mich als Vorstellenden, der ich am Schreibtisch sitze, nicht emotional berühren.[2]

II.

Nachdem ich soeben Modi unterschieden habe, *wie* ich mir etwas möglicherweise vorstellen kann, werde ich nun in den folgenden vier Abschnitten Möglichkeiten durchgehen, *was* ich dabei versuchen könnte mir vorzustellen. In der Unterscheidung der ersten beiden Möglichkeiten folge ich dabei in etwa Goldie (vgl. Goldie 2000, 194–205).

Zuerst *In-deinen-Schuhen-Stecken*. Ein klassischer Fall hiervon wäre folgender: Ich stelle mir vor, dass ich durch irgendeine Zauberei plötzlich im Körper von jemand anderem stecke. Dabei bleiben äußere physische Eigenschaften der Person erhalten. Ich kann mir hier vielleicht außerdem noch vorstellen, dass ich über ihr Wissen oder ihre (Quasi-)Erinnerungen verfüge. Trotzdem bin ich es, der Entscheidungen fällt. Ich „stecke in ihren Schuhen", kann mich dabei aber bewusst anders verhalten, als sie es tun würde.

2. Sie kann mich berühren, aber nicht als Teil der Vorstellung, sondern durch einen nachgeordneten Prozess, der bei Goldie „emotionale Ansteckung" heißt.

Besser gesagt bleibt hier mein Selbst erhalten und wird in die andere Person „verpflanzt". Den Begriff „Selbst" möchte ich bewusst vage lassen, denn es gibt hier viele Möglichkeiten, die Rahmenbedingungen zu variieren. Wenn ich mir vorstellen möchte, ich steckte in den Schuhen einer depressiven Person, dann kann ich mir mich als Person ohne Depressionen, in ihrer Lebenssituation (und ihrem Körper usw.) vorstellen, oder eine depressive Variante von mir in ihr handeln lassen.

Wenn ich mir also in diesem Sinne vorstelle, ich sei Joachim Gauck, dann stelle ich mir nicht nur vor, dass ich viel herumreisen und Pressekonferenzen geben müsste, sondern ich überlege auch, was ich an seiner Stelle anders oder auch genauso machen würde. Ich kann mir nun in allen drei Modi vorstellen, ich steckte in Gaucks Schuhen: Erstens ich-zentriert, indem ich mir überlege wie ich, mit den äußeren Eigenschaften des echten Gauck ausgestattet, als Akteur die Rolle des Bundespräsidenten so ausfülle, wie ich es möchte, und mir die „innere" emotionale Situation aneigne, in die ich dabei gerate. Oder zweitens und drittens azentriert oder peripher, indem ich mir ausmale, wie ich aus der „Vogelperspektive" beobachte, oder als Außenstehender aus den Nachrichten erfahre, wie Gauck als Person mit meinen Charakterzügen agiert.

Alle drei Typen von imaginativen Projekten lassen sich umso besser verwirklichen, je besser ich über die andere Person informiert bin. Und auch ein klares Konzept davon, was mich selbst ausmacht, hilft mir dabei. Die Unterschiede zwischen den Realisierungen solcher Projekte in Bezug auf verschiedene „Zielpersonen" sowie verschiedene ausführende Personen scheinen mir jedoch nur graduell zu sein. Das wird auch daran deutlich, dass ich mir – wie auch Goldie – vorstellen kann, ein römischer Legionär zu sein, der auf einer antiken Wanderstraße in der Hitze entlangwandert, auch ohne irgendetwas Konkretes über diesen Soldaten zu wissen, das ihn von seinen Kameraden unterscheidet.[3]

In-deinen-Schuhen-Stecken ist sicherlich eine gängige Reaktion auf die Aufforderung, sich vorzustellen, eine andere Person zu sein. Die Frage „Stell dir vor

3. Vgl. Goldie 2000, 204. Dieser Punkt steht jedoch in einer Spannung zu Goldies Verständnis von „in-his-shoes-imagining", das Fälle ausschließt, in denen ich nichts über die Person selbst weiß, sondern lediglich ein wenig über die Situation, in der sie sich befindet, weil sich das Ergebnis des Vorstellungsprozesses nicht von dem Ergebnis anderer imaginativer Prozesse unterscheiden lässt (denn wir gingen hier jeweils von unserem eigenen Charakter aus, vgl. Goldie 2000, 200f.). Allerdings denke ich, dass sich hier begriffliche Eindeutigkeit besser erreichen lässt, indem wir im nächsten Abschnitt für den Begriff der Einfühlung voraussetzen, dass tatsächlich fremder Charakter übernommen wird, und den bei Goldie eigenständigen Fall des Selbst-in-einer-Situation-Handelns als Spezialfall von In-deinen-Schuhen-Stecken auffassen (das ist konsistent mit Obigem). So vermeiden wir ein Vagheitsproblem in Bezug auf die Frage, wie viel ich über jemanden wissen muss, um mir vorstellen zu können, dass ich in ihren oder seinen Schuhen stecke.

du wärst Kennedy; wie würdest du dich in der Kuba-Krise verhalten?“ fordert diese Reaktion heraus und scheint gleichzeitig völlig unproblematisch zu sein.

Die imaginativen Projekte des In-deinen-Schuhen-Steckens enthalten keine Identität von mir und jemand anderem. Vielmehr stelle ich mir eine Art Mischung aus mir und der anderen Person vor, die von mindestens einem von uns verschieden ist. Die vorgestellte Person hat etwa meinen Charakter, aber die äußeren Merkmale der anderen Person. Egal was wir als Kriterium der Personenidentität annehmen, die vorgestellte Person kann höchstens mit einer der Ausgangspersonen identisch sein, da sie kein Merkmal aufweist, das auf beide Ausgangspersonen verweist.

III.

Nun komme ich zum zweiten Typ imaginativer Projekte, dem *Einfühlen*. Mit Einfühlen meine ich hier affektive Empathie (im Gegensatz zu kognitiver Empathie, siehe Walter 2012) – ein Projekttyp, in dem ich versuche die Erfahrungen der Anderen zu spiegeln. Wenn ich mich in eine Person einfühle, bin ich gewissermaßen nur Mitfahrer: Ich versuche auf Grundlage meines Verständnisses der Person ihre Gedanken, Wahrnehmungen, Gefühle und Wünsche im Rahmen ihrer Persönlichkeit nachzuerleben oder eine ebenfalls von mir ausgedachte Situation so zu erleben, wie diese Person sie erleben würde. Ich könnte so auch versuchen, mich in eine bestimmte Person in einer möglichen zukünftigen Situation einzufühlen, um ihr Verhalten in eben dieser Situation vorhersagen zu können.[4]

Dass Einfühlen von In-deinen-Schuhen-Stecken grundsätzlich verschieden ist, zeigt sich in folgendem Beispiel: Anna und Ben stellen sich jeweils vor, sie wären Caro. Zuerst stellen sie sich vor, sie selbst steckten jeweils in Caros Schuhen. Dabei kommen sie zu unterschiedlichen Ergebnissen – was aber nicht bedeutet, dass ihre Vorstellungen unterschiedlich gute Verwirklichungen ihrer imaginativen Projekte sind. Schließlich wollte sich Anna vorstellen, was wäre,

4. Allerdings ist nicht jede Vorhersage durch Einfühlung zustande gekommen: Wenn ich weiß, dass Anna Höhenangst hat, und daraus ableite, dass sie nicht auf den Turm möchte, ist das keine Einfühlung, sondern rein theoretische Reflexion über Anna. Diese Reflexion wäre wohl besser als kognitive Empathie zu bezeichnen (vgl. Walter 2012, 10). Zwar ist es sicher angebracht, auch diese mentalen Vorgänge Empathie (jedenfalls in einem weiten Sinne) zu nennen, allerdings denke ich, dass es sich dabei nicht mehr um eine Vorstellung (im Sinne eines imaginativen Projekts) handelt – sondern um eine rein theoretische Leistung, der jeglicher Erlebnischarakter fehlt. Sicherlich gibt es auch Leute, die über nicht erlebnisartige Vorstellungen reden – im Sinne eines bloßen „Vor-sich-hin-Stellens“ –, ich möchte mich hier aber auf Erlebnis-Vorstellungen beschränken.

wenn sie, Anna, an Caros Stelle stünde, und Ben wollte sich vorstellen, was wäre, wenn er, Ben, in Caros Lage wäre. Sie hatten also zwei unterschiedliche imaginative Projekte.[5] Anschließend versuchen nun beide sich vorzustellen, was in Caro selbst vorgeht und wie es für sie ist, ihr Leben zu leben. Wenn wir annehmen, dass Anna und Ben Caro beide sehr gut kennen und beide ähnlich einfühlsam sind, dann sollten sich ihre Vorstellungen auch sehr ähneln. Die beiden haben hier das gleiche imaginative Projekt. Während also beim In-deinen-Schuhen-Stecken zwei Personen Teil des imaginativen Projekts sind, ist es beim Einfühlen nur eine.

Noch ein weiterer wichtiger Unterschied besteht zwischen diesen beiden Vorhaben: Einfühlen ist anders als In-deinen-Schuhen-Stecken nur als ich-zentriertes imaginatives Projekt sinnvoll. Wenn ich mir nur azentriert oder peripher vorstelle, wie eine bestimmte Person in irgendeiner Situation handelt (auch wenn sie dabei ihr Gefühlsleben etc. zu erkennen gibt), beteilige ich mich dabei zunächst nicht an ihren Emotionen; von Einfühlen kann in solch einem Fall nicht die Rede sein.

Inwieweit kann ich mich überhaupt in eine andere Person einfühlen? „You don't know how it feels to be me", singt Tom Petty. Sicherlich ist es häufig so, dass wir vieles über die andere Person nicht wissen und eine schlicht falsche Vorstellung davon haben, wie es sich anfühlt, z.B. Tom Petty zu sein. Diese Schwierigkeiten lassen sich aber prinzipiell und zumindest in einigen Fällen überwinden – sie beruhen auf einem Informationsdefizit oder vielleicht einem Mangel an Einfühlungsvermögen bzw. eigenen, hinreichend ähnlichen Erfahrungen. Es gibt aber auch häufig *zumindest bis zu einem gewissen Grad* geglückte Umsetzungen imaginativer Projekte des Einfühlens.

Nicht Teil dieser Projekte ist aber sicherlich die Berücksichtigung von Unterschieden in Bezug auf Qualia.[6] Dass wir immer wieder versuchen, uns in nahestehende Personen einzufühlen, zeigt doch, dass es uns dabei nicht um die Überwindung dieser doch per se unüberwindbaren qualitativen Barriere zu Anderen geht, sondern um eine möglichst gute Annäherung unter der Annahme, dass die andere Person phänomenal auf eine ähnliche Weise Dinge erlebt wie wir.

Einfühlung ist sicherlich ein weiteres übliches Verständnis von „sich vorstellen, eine andere Person zu sein". Aber auch hier wird „sein" nicht als „identisch sein" verstanden. Wenn ich versuche, die realen oder hypothetischen mentalen Ereignisse einer anderen Person mitzuerleben, lege ich meine eigene Persönlichkeit für einen Moment beiseite und konzentriere mich nur auf die andere. Die betreffende Person bleibt dabei von mir klar getrennt, sodass ich mir vorstelle,

5. Natürlich hätte sich Ben auch vorstellen können, was wäre, wenn Anna in Caros Schuhen steckte. Wenn wir etwa annehmen, beide hätten sich azentriert vorgestellt, Anna steckte in Caros Schuhen, dann hätten sie durchaus das gleiche imaginative Projekt.

6. Die gibt es übrigens wirklich.

wie es ist (also: sich anfühlt) eine andere Person zu sein, aber nicht, dass ich mit ihr identisch bin.

IV.

Die bisher untersuchten imaginativen Projekte enthielten nicht, dass ich mir vorstelle, mit jemandem *identisch* zu sein, sondern mich der Person lediglich nähere. Liegt das womöglich daran, dass es unmöglich ist, mit einer anderen Person identisch zu sein? David Lewis scheint das zu meinen:

> […] Heimson couldn't be Hume. If he believes the proposition that holds at just those worlds where he is Hume, then he believes the empty proposition that holds at no worlds. In the first place, there is no world where Heimson and Hume are literally identical. Suppose there were; then from the standpoint of that world, their difference at this world would be a difference between Hume and Hume, which is absurd. (Lewis 1979, 524)[7]

Hier ist Lewis auf einer Linie mit Saul Kripke, der argumentiert, dass Identitätsaussagen mithilfe von Eigennamen metaphysisch notwendig wahr oder falsch sind (vgl. Kripke 1980, 97–105). Aber andererseits kann ich auch metaphysisch Unmögliches glauben, was Kripke in seiner Unterscheidung von *metaphysischer* und *epistemischer* (i.e. apriorischer) *Notwendigkeit und Möglichkeit* ausdrücklich zulässt (vgl. Kripke 1980, 34–38). Dem Verständnis dieser Unterscheidung von David Chalmers nach können wir zwischen dem Vorstellen von Szenarien (die epistemischen Möglichkeiten entsprechen) *als aktual* und dem *kontrafaktischen* Vorstellen von metaphysisch möglichen Welten unterscheiden (vgl. Chalmers 2002, 609b–611a).

Ein Vorschlag wie ich mir, die epistemische Möglichkeit ausnutzend, vorstellen kann, mit Napoleon identisch zu sein, ist folgender:

> As Eric Lormand has pointed out to me, however, there are many ways to imagine that I am Napoleon, including [...] imagining that Napoleon has been reincarnated as David Velleman, or that he was cryogenically preserved at birth, thawed out in 1952, and handed by the maternity nurses to an unsuspecting Mrs. Velleman. (Velleman 2006, 171, Fn. 2)

Der erste Vorschlag einer Wiedergeburt ist aus meiner Sicht einleuchtend, allerdings gibt es sicher auch Leute, die dazu sagen würden: „Neues Leben, neue Identität." Dieses Problem hat der zweite Vorschlag nicht – aber vielleicht ist

7. Hier geht es um „glauben", jedoch ist das Problem auch auf azentriertes oder peripheres Vorstellen anwendbar. Ich-zentriertes Vorstellen ist für diese Strategie nicht sinnvoll, siehe unten.

seine Schwäche, dass das eingefrorene Kleinkind nicht der Napoleon ist, den wir typischerweise meinen, wenn wir über „Napoleon" reden (denn wir meinen den Napoleon, der in Waterloo war). Folgende Variation umgeht das: Ich stelle mir vor, dass 1769 ein Baby auf den Namen „Napoleon" getauft wurde und anschließend die bekannte Napoleon-Biografie durchlebt. Auf St. Helena wurde dessen Tod nur vorgetäuscht, Napoleon wird in Wahrheit jedoch chirurgisch in den Zustand eines Babys zurückversetzt und seine Erinnerungen werden gelöscht. Dann wird er bis 1989 eingefroren und wie oben beschrieben im Krankenhaus heimlich eingetauscht, sodass ich Napoleon wäre.[8]

In Kripkes Bild der Referenz von Eigennamen ist kein Hindernis angelegt, dieses Beispiel zu akzeptieren: Wenn wir einen Namen verwenden, dann referiert er auf denjenigen Gegenstand, dessen Taufakt am Anfang einer Kausalkette steht, entlang der der Name weiter verbreitet wird und an deren Ende meine Bekanntschaft mit dem Namen steht (vgl. Kripke 1980, 91f.). Dann aber ist der Referent in den obigen Beispielen derselbe.

Ich kann mir solche Szenarien azentriert oder peripher vorstellen, indem ich mir vorstelle, den Verlauf des Geschehens zu beobachten. Bei einem ich-zentrierten imaginativen Projekt stehe ich vor dem Problem, mir die Erklärung vorzustellen, warum ich mit Napoleon identisch bin. Ich kann mir vorstellen, dass meine Erinnerungen nur schlampig gelöscht wurden und ich mich nun langsam an meine Napoleon-Vergangenheit und den „Transformationsprozess" erinnere. Aber diese Vorstellung ist von einer ich-zentrierten Vorstellung davon, dass ich verrückt werde, nicht zu unterscheiden.

V.

Der in IV. beschriebene imaginative Projekttyp lässt sich durchführen, weil es epistemisch möglich ist, dass ich tatsächlich mit einer Person identisch bin, von der ich meine, dass sie von mir verschieden ist. Nun könnte aber – wenn wir diese Phrase ganz wörtlich nehmen wollen – noch mehr von der Vorstellung meiner Identität mit einer *anderen* Person gefordert werden: Nämlich, dass ich mir *kontrafaktisch* vorstelle, mit einer Person identisch zu sein, und dabei gleichzeitig festsetze, dass diese Person in Wirklichkeit nicht mit mir identisch ist.

Auf den ersten Blick scheint es, dieser Aufgabenstellung könnten wir fast genauso begegnen wie oben: Angenommen ich bin nicht identisch mit Napoleon.

8. Ob diese Beispiele als Fälle von vorgestellter Personenidentität akzeptiert werden, hängt allerdings von den Anforderungen hierfür ab. Wenn wir physische Kontinuität fordern oder eine Seele bzw. cartesisches Ego zur Voraussetzung für Identität machen, dann ja. Kontinuität von Erinnerungsketten oder Persönlichkeitsmerkmalen liegt dagegen nicht vor.

Stellen wir uns nun (azentriert) eine Situation vor, in der Napoleon wie oben behandelt und 1989 heimlich durch einen Babytausch meiner Mutter untergeschoben wird. Aber was haben wir uns nun vorgestellt? Dass ich als Kleinkind im Krankenhaus vertauscht und durch Napoleon ersetzt wurde. In dieser Vorstellung bin ich nicht Napoleon, der meiner Mutter untergeschoben wird. Ich bin das ausgetauschte Kind.

Wir können uns auch vorstellen, dass ich Napoleons Eltern untergeschoben werde. Oder wir stellen uns eine Welt vor, in der Napoleon oder ich nicht existieren und die entstehende „Lücke" durch den jeweils anderen gefüllt wird. Aber in keiner dieser Vorstellungen sind Napoleon und ich identisch – hier greift Lewis' Argument von oben. Eine Vereinigung von Napoleon und mir kann ich mir nur vorstellen, wenn ich die Annahme infrage stelle, dass wir beide in Wirklichkeit verschiedene Personen sind. Alle anderen imaginativen Projekte scheitern – im Grunde an genau dem Problem, das nach Kripke nicht-aktuale Identitätsbeziehungen metaphysisch unmöglich macht.

VI.

Hier ein Überblick über die Vorstellbarkeit der zwölf systematisch denkbaren imaginativen Projekttypen:

	In-deinen-Schuhen-Stecken	Einfühlen	Identität in gerader Linie	Vereinigung
ich-zentriert	✓	✓	/	×
azentriert	✓	/	✓	×
peripher	✓	/	✓	×

Möglichkeit imaginativer Projekte (Haken: möglich, Strich: nicht sinnvoll, Kreuz: scheitert)

Die „nicht sinnvollen" Projekttypen repräsentieren das gewünschte Szenario nicht angemessen. Diese Projekte sind nicht inkohärent, aber wenn ich sie durchführe, werde ich hinterher nicht von mir behaupten können, ich hätte mir vorgestellt, eine andere Person zu sein. Bei den Projekten mit einer Vereinigung stoße ich dagegen auf eine nicht überwindbare gedankliche Barriere.

Halten wir fest: In den gängigsten Verständnissen von „eine andere Person sein" kann ich mir etwas darunter vorstellen. Wenn wir das „sein" als Identität verstehen, wird es schwieriger: Zwar kann ich die epistemische Möglichkeit ausnutzen, dass ich tatsächlich mit einer beliebigen Person identisch sein könnte, aber dann kann ich eigentlich nicht sagen, dies sei eine *andere* Person. Aber wenn

ich voraussetze, dass ich mit einer bestimmten anderen Person in Wirklichkeit nicht identisch bin, kann ich mir nicht mehr vorstellen, mit ihr identisch zu sein.[9]

Literatur

Chalmers, David 2002: "The Components of Content". In: Ders. (Hg.), *Philosophy of Mind. Classical and Contemporary Readings*. Oxford: Oxford University Press, 608–633.

Goldie, Peter 2000: *The Emotions. A Philosophical Exploration*. Oxford: Clarendon Press.

Kripke, Saul 1980: *Naming and Necessity.* Oxford: Blackwell.

Lewis, David 1979: "Attitudes de dicto and de se". *The Philosophical Review* 88, 513–543.

Velleman, David 2006: "Self to Self". In: Ders. (Hg.), *Self to Self. Selected Essays*. Cambridge: Cambridge University Press, 170–202.

Walter, Henrik 2012: "Social Cognitive Neuroscience of Empathy: Concepts, Circuits and Genes". *Emotion Review* 4, 9–17.

Williams, Bernard 1973: "Imagination and the Self". In: Ders. (Hg.), *Problems of the Self. Philosophical Papers 1956–1972*. Cambridge: Cambridge University Press, 26–45.

Wollheim, Richard 1984: *The Thread of Life*. Cambridge, Mass.: Harvard University Press.

9. Ich bedanke mich für hilfreiche Kommentare und Anregungen zu früheren Versionen dieses Textes bei Sören Hilbrich, dem Oberseminar von Simon Friederich (mit Johanna Mardt, Mark Thomsen, Julian D. Small und Anton Alexandrov), der Jury der Preisfrage und Achim Stephan. Den Hinweis auf Tom Petty verdanke ich Timm Fitschen.

Grazer Philosophische Studien
90 (2014), 317–326.

ESSAY ZUR FRAGE: *KANN ICH MIR VORSTELLEN, EINE ANDERE PERSON ZU SEIN?*

Eva Backhaus
Goethe-Universität Frankfurt/M.

Gewinnerin des zweiten Preises des Essay-Wettbewerbs für Studierende 2013
gesponsert durch die Gesellschaft für Analytische Philosophie (GAP)
in Kooperation mit den Grazer Philosophischen Studien*

Summary
This essay is concerned with the question whether we can imagine being another person. I argue that the answer is negative because it is both logically impossible to become another person and impossible to imagine a logical impossibility. To imagine a state of affairs is to know its truth conditions, and logical impossible states of affairs do not have truth conditions; therefore it is impossible to imagine them. Luckily the epistemic and aesthetic merits which stem from the apparent possibility to imagine being another person can be much better accounted for by imagining being *like* another person.

Eine andere Person sein

Spontan würden die meisten Menschen die Frage, ob sie sich vorstellen können, eine andere Person zu sein, mit einem deutlichen „Ja" beantworten. Und das ist nicht erstaunlich, denn wir sprechen häufig so, dass wir „uns in jemanden hineinversetzen", „es aus seiner Sicht sehen", oder werden aufgefordert, uns vorzustellen, wir wären an der Stelle von jemand anderem. Empathische Personen können das – ihnen fällt es leicht nachzuvollziehen, wie die Welt aus den Augen anderer aussieht. Häufig gehen uns die Schicksale anderer Personen – seien sie

* Die Preisfrage des Wettbewerbs 2013 war: „Kann ich mir vorstellen, eine andere Person zu sein?" Unter den 36 Einreichungen, die den Regeln des Wettbewerbs entsprachen, wählte die Jury drei Beiträge aus, die den jeweils ersten, zweiten und dritten Platz belegten. Die Autorinnen und Autoren der Gewinnerbeiträge erhielten die Erlaubnis, ihre Essays für die endgültige Publikation geringfügig zu überarbeiten.

real oder fiktiv – so nahe, dass wir mitfiebern, mitleiden und uns mit ihnen freuen. Im Film und in der Literatur gibt es Kameraführungen und Erzählweisen, die uns eine Sicht auf die Welt präsentieren, die nicht die unsere ist. Wenn wir mit Emma Bovary die dicken Finger des armen Charles betrachten und uns eines Gefühls der Verachtung nicht erwehren können, dann doch deshalb, könnte man meinen, weil wir uns vorstellen, Emma zu sein: zu sehen, was sie sieht, zu fühlen, was sie fühlt, und zu wissen, was sie weiß.

Trotz der überwältigenden Zahl an vermeintlichen Gegenbeispielen, die unsere Bücherregale und DVD-Sammlungen bewohnen, behaupte ich, dass niemand sich vorstellen kann, eine andere Person zu sein als die Person, die er nun mal ist, und zwar aus einem einfachen Grund: Dass ich oder irgendjemand eine andere Person werden kann, ist metaphysisch unmöglich, da es sich bei Personen um Einzeldinge handelt.[1] Wenn wir uns fragen, ob eine Person zu einer anderen Person werden kann, die sie jetzt noch nicht ist, fragen wir also, ob ein bestimmtes Einzelding (Person A) zu einem anderen Einzelding werden kann (Person B). Identität ist eine transitive Beziehung; dies bedeutet, dass wenn Person B zu t_2 mit A zu t_2 identisch ist und A sowohl zu t_1 als auch zu t_2 mit sich selbst identisch ist, Person B bereits zu t_1 mit A identisch sein müsste. Nach Voraussetzung sollen Person A und B zu t_1 aber nicht identisch sein, da es sich um unterschiedliche Personen und damit um unterschiedliche Einzeldinge handelt. Diese Überlegungen zeigen, dass kein Einzelding zu einem anderen Einzelding werden kann und somit niemand zu einer anderen Person werden kann. Transitivität ist eine der formalen Eigenschaften der numerischen Identität von Einzeldingen, die besagt, dass jedes Einzelding mit sich selbst und nur mit sich selbst identisch ist. Da es sich bei Personen um Einzeldinge handelt, ist ausgeschlossen, dass eine Person eine andere Person ist oder werden kann. Eine andere Person zu werden, stellt eine metaphysische

1. Manche werden sich an der Formulierung, dass Personen Einzeldinge sind, stören. Mit dieser Aussage ist jedoch nicht gemeint, dass man über Personen (oder das Wesen von Personen) nicht mehr oder Interessanteres sagen kann, als dass sie Einzeldinge sind. Doch davon bleibt die Tatsache unberührt, dass Personen in ontologischer Hinsicht Einzeldinge sind und die Frage danach, ob wir eine andere Person werden können, somit bedeutet, dass wir fragen, ob wir zu einem anderen Einzelding werden können. Dieser Sinn des Begriffs der Person ist natürlich ein anderer, als wenn wir ihn benutzen, um den Charakter einer Person zu beschreiben, wie es in Aussagen wie „Seit ihrer Scheidung ist sie eine ganz andere Person" geschieht. Mit solchen Bemerkungen wollen wir natürlich nicht sagen, dass es eine Person gab, die es nach der Scheidung nicht mehr gibt, und dafür eine buchstäblich andere Person begonnen hat zu existieren. Viel eher handelt es sich um ein und dieselbe Person, die nach ihrer Scheidung ihren Charakter (möglicherweise in umfassender Weise) geändert hat. Insofern kommen sowohl die Charaktereigenschaften vor der Scheidung als auch die Charaktereigenschaften nach der Scheidung derselben Person zu. Und um diesen Sinn des Begriffs der Person geht es mir hier.

Unmöglichkeit dar, die in der logischen Unmöglichkeit gründet, eine andere Person (d.h. ein anderes Einzelding) zu sein als die Person, die man nun einmal ist.

Nun könnte man an dieser Stelle einwenden, dass es zwar logisch unmöglich ist, eine andere Person *zu sein*, und daher unmöglich*, die Überzeugung zu haben*, dass man eine andere Person ist oder werden könnte, es aber nicht unmöglich ist, sich *vorzustellen*, eine andere Person zu sein. Es macht ja gerade den Witz von Vorstellungen aus, dass sie sich nicht danach richten müssen, was tatsächlich der Fall ist oder sein könnte, sondern höchstens danach, was *innerhalb der Vorstellung* der Fall ist oder sein könnte.

Arten von Vorstellungen

Eine solche Position wird unter anderem von Colin McGinn in dem Buch *Mindsight* (2004) vertreten. Er unterscheidet zwischen zwei verschiedenen Arten von Vorstellungen: den sensorischen und den kognitiven Vorstellungen. Erstere sind auditive, visuelle oder taktile Vorstellungen, die auf neue Arten zusammengesetzt werden können, aber es nicht sein müssen (Erinnerungen haben zumindest den Anspruch, auf dieselbe Weise zusammengesetzt zu sein wie in der zurückliegenden Wahrnehmungssituation). Nach dem gleichen Prinzip und vom selben Vermögen ausgeführt sollen auch kognitive Vorstellungen verstanden werden. Eine kognitive Vorstellung besteht darin, dass wir uns einen bestimmten Sachverhalt vorstellen, wie z. B. „Ich stelle mir vor, dass ich in Paris bin" oder „Ich stelle mir vor, dass da ein Tausendeck ist". Kognitive Vorstellungen weisen keinen Erlebnisaspekt auf, sondern können höchstens durch sensorische Vorstellungen ergänzt werden (wie es wahrscheinlich im ersten, aber eher nicht im zweiten Beispiel der Fall ist). Wenn man davon ausgeht, dass „sich etwas vorstellen" bedeutet, dass man sich vorstellt, wie man dieses oder jenes erleben oder empfinden würde, könnte man meinen, es handele sich bei kognitiven Vorstellungen gar nicht um richtige Vorstellungen. Doch selbst wenn man bestreitet, dass es sich bei kognitiven Vorstellungen um Vorstellungen in einem interessanten Sinn handelt, stellen sie wichtige Vorbedingungen für umfassendere Vorstellungen dar. Wenn ich mir vorstelle, wie ich auf das Meer blicke, stelle ich mir vor, wie es sich anfühlt, wenn mir der Wind ins Gesicht weht oder wie die Luft schmeckt, in einem grundlegenden Sinn bedeutet es aber auch, dass ich weiß, was in der Welt der Fall sein müsste, damit der Satz, der den basalen Gehalt der Vorstellung ausmacht („Ich blicke aufs Meer"), wahr ist. Diese Tatsache mag uns häufig entgehen, da wir uns vor allem für reichhaltigere, sensorische Vorstellungen interessieren. Aber auch diese können uns, indem

wir uns vorstellen, *wie* etwas ist, darauf festlegen uns vorzustellen, dass *etwas der Fall* ist.[2]

Kann man sich logische Unmöglichkeiten vorstellen?

Eine zutreffende Beobachtung McGinns ist, dass Vorstellungen im Unterschied zu Überzeugungen nicht gegenüber der Welt verantwortlich sind. Wie die Welt beschaffen ist, sollte zwar den Maßstab für meine Überzeugungen darstellen, Vorstellungen hingegen zeichnen sich gerade dadurch aus, dass sie gegenüber der Wahrheit indifferent sind. Man kann sich beispielsweise im vollen Bewusstsein davon, dass der Eiffelturm eine graue Farbe hat, vorstellen, er sei gelb. Die Überzeugung „p" ist zwar unvereinbar mit der *Überzeugung* „nicht p", aber vereinbar mit der *Vorstellung* „nicht p". Und dies scheint ein richtiger Gedanke zu sein, da „sich etwas vorstellen" ja gerade darin besteht, sich etwas zu vergegenwärtigen, das entweder in der Situation, in der man sich gerade befindet, oder in der Welt überhaupt nicht vorhanden ist. Aus dieser Beobachtung folgert McGinn, dass Vorstellungen keine Überzeugungen sind, denn Überzeugungen beinhalten, dass man sich darauf festlegt, ob „p" besteht oder nicht besteht. Wenn wir uns einen Sachverhalt *vorstellen*, müssen wir jedoch weder davon überzeugt sein, dass der Sachverhalt besteht, noch dass der Sachverhalt möglich ist. Vielmehr verhalten wir uns gewissermaßen qua Vorstellendem neutral zum modalen Status des Sachverhalts. Jeder grammatische Satz soll dazu in der Lage sein, einen Sachverhalt zu repräsentieren, dessen modaler Status zunächst nicht relevant ist und höchstens in einem zweiten Schritt geprüft werden kann (vgl. McGinn 2004, 138, 155f). McGinn trennt also die Vorstellbarkeit eines Sachverhalts von dessen Möglichkeit in strikter Weise. Doch die Beobachtung, dass viele Vorstellungen Sachverhalte repräsentieren, ohne sich um das tatsächliche Bestehen des Sachverhalts oder die naturgesetzliche Möglichkeit desselben scheren zu müssen, lässt, behaupte ich, die Tatsache unberührt, dass es nicht möglich ist, einen Sachverhalt zu *repräsentieren*, wenn der Sachverhalt eine logische oder metaphysische Unmöglichkeit darstellt.

Um zu wissen, was wir uns unter einem Satz vorstellen sollen, müssen wir wissen, welchen Sachverhalt der Satz repräsentiert. Und dies bedeutet, dass wir wissen, was der Fall sein müsste, damit dieser Satz wahr wäre, d.h. wir müssen

2. Das bedeutet jedoch nicht, dass allen sensorischen Vorstellungen eine kognitive Vorstellung zugrunde liegt. Das Im-Geiste-„Singen" einer Melodie, die Vorstellung des Geruches von Bienenwachs oder die Vorstellung eines leuchtenden Gelbs setzen nicht voraus, dass man sich vorstellt, dass diese Musik gerade gespielt wird, oder dass man sich vor einer Bienenwabe befindet.

seine Wahrheitsbedingungen kennen.[3] Eine solche Beschreibung des Zusammenhangs zwischen Vorstellbarkeit und Wahrheitsbedingungen wird unter anderem von Stephen Yablo vorgeschlagen: „*P* is conceivable for me if *I* can imagine a world that I take to verify *P*." (Yablo 1993, 29) Dass wir wissen, was der Fall sein müsste, damit ein beliebiger Satz wahr ist, kann auch in Fällen gegeben sein, in denen wir uns das Gegenteil einer wahren Überzeugung vorstellen. (Da wir angeben, was der Fall sein *müsste* und nicht was der Fall *ist*, stehen die Vorstellung und die ihr widersprechende Überzeugung tatsächlich nicht in Konkurrenz.) In dem oben genannten Beispiel können wir leicht angeben, wie die Welt beschaffen sein müsste, damit der Satz „Der Eiffelturm ist gelb" wahr wäre, nämlich so, dass der Eiffelturm gelb ist. Und so geht es auch in Fällen, die uns selbst betreffen. Wenn ich mir vorstelle, ein Mann, Metzger oder zwei Meter groß zu sein, kann ich recht genau angeben, was in der Welt der Fall sein müsste, damit die entsprechenden Sätze wahr wären.

Nicht möglich ist es hingegen, sich logisch Unmögliches vorzustellen, da logische Unmöglichkeiten gerade dadurch gekennzeichnet sind, dass es keine Wahrheitsbedingungen für sie gibt. Im Fall einer logischen Unmöglichkeit können wir nicht angeben, was der Fall sein müsste, damit der entsprechende Satz wahr wäre. Wenn wir versuchen, uns eine logische Unmöglichkeit wie ein rundes Viereck vorzustellen, würde diese Vorstellung besagen, dass es ein Objekt gibt, das eine Eigenschaft besitzt (die Rundheit) und zugleich dieselbe Eigenschaft nicht besitzt (denn Viereckigkeit schließt Rundheit aus). Und wir können uns keine Situation vorstellen, in der das Urteil „x ist viereckig" und das Urteil „x ist nicht viereckig" zugleich wahr sind.

Schwache und starke Vorstellungen

McGinn behauptet hingegen, dass wir uns logisch Unmögliches in einem *schwachen* Sinn vorstellen können, denn für kognitive Vorstellungen soll es ausreichen, dass sie sich in einem grammatischen Satz ausdrücken lassen – und der Satz „x ist ein rundes Viereck" ist durchaus grammatisch. Doch was könnte damit gemeint sein, dass man sich den entsprechenden Sachverhalt in einem schwachen Sinn vorstellen kann? McGinn behauptet, dass wir auch in solchen Fällen wissen „what it is we are supposed to be getting our mind around." (McGinn 2004, 155) Auch

3. Eine Tatsache, die auch McGinn für richtig hält. Ich teile daher seine Überlegungen zu Vorstellungen von tatsächlich bestehenden Sachverhalten und zu Vorstellungen, die nicht-aktuale oder empirisch nicht mögliche Sachverhalte betreffen. Unsere Wege trennen sich jedoch, wenn es um logische oder metaphysische Unmöglichkeiten bzw. *necessary falsehoods* geht (vgl. McGinn 2004, 155–158).

Chalmers sieht, dass es Fälle gibt, in denen jemand behaupten könnte, dass er sich logische Unmöglichkeiten vorstellen kann. Eine Person, die so etwas behauptet, „imagines a situation in something less than full detail." (Chalmers 2002, 152) Dass Leute meinen, sie könnten sich eine Situation vorstellen, die logische Unmöglichkeiten beinhaltet, hieße dann, dass sie sich die Situation auf eine solch rudimentäre Weise vorstellen, dass ihnen die Widersprüchlichkeit nicht auffällt. Vorstellungen können, im Gegensatz zu Wahrnehmungen, unvollständig sein, um sich etwas vorzustellen, muss man nicht, und kann man auch gar nicht, jedes Detail imaginieren. Doch die Vorstellung muss zumindest so reichhaltig sein, dass man sich die relevanten Details vorstellt und in der Lage wäre, beliebig viele Details, die aus den gemachten Annahmen folgen, widerspruchsfrei einzufügen. Chalmers definiert Vorstellbarkeit einer Proposition daher folgendermaßen: „S is positively conceivable when one can coherently modally imagine a situation that verifies S." (Chalmers 2002, 153) Dabei bedeutet „coherently modally imaginable", dass die Vorstellung nicht widersprüchlich sein oder Widersprüchliches aus ihr folgen darf. Folglich kann man sich bei einer Vorstellung zunächst über ihre Widerspruchsfreiheit täuschen, müsste diese Vorstellung aber zurückweisen, wenn man entweder aufgefordert würde, Details zu ergänzen, und sich daraus ein Widerspruch ergäbe, oder auf die widersprüchlichen Implikationen der Annahmen hingewiesen würde.

Die Vorstellung, man sei buchstäblich eine andere Person, ist widersprüchlich, da man sich vorstellen müsste, dass man ein bestimmtes Einzelding (Person A) ist, und zugleich, dass man ein anderes Einzelding (Person B) ist, und kein Einzelding kann zugleich ein anderes (numerisch verschiedenes) Einzelding sein. Daher kann man sich nicht auf kohärente Weise eine Situation vorstellen, in der man eine andere Person ist als man selbst.

Wie kann man sich vorstellen, eine andere Person zu sein?

Trotz alledem finden wir es nicht unverständlich, wenn jemand davon spricht, er stelle sich vor, eine andere Person zu sein. In diesem Abschnitt möchte ich erläutern, was damit gemeint sein kann, sich vorzustellen, man sei eine andere Person.

Man kann sich sicherlich vorstellen, man habe verschiedene Eigenschaften einer anderen Person oder bestimmte seiner aktualen Eigenschaften nicht. Solche Vorstellungen können darin bestehen, dass man lediglich eine Eigenschaft in der Vorstellung verändert, aber natürlich kann man sich auch vorstellen, dass man zahlreiche Eigenschaften einer anderen Person hat und fast keine seiner eigenen, aktualen Eigenschaften. Das ändert aber nichts daran, dass man sich

damit jeweils etwas *über sich selbst* vorstellt und daher stellt man sich in einem solchen Fall genau genommen nicht vor, *eine andere Person zu sein*, sondern in vielen Hinsichten *wie eine andere Person zu sein*.

Nun könnte man an dieser Stelle überlegen, ob man sich nicht in einem Vorstellungsakt eine andere Person vorstellen kann und in einem zweiten davon unabhängigen Vorstellungsakt, dass man selbst mit dieser Person identisch ist. So hätte man die oben diskutierte Widersprüchlichkeit in der Vorstellung umschifft, da man in der Vorstellung einer anderen Person, die man sich ja zunächst nicht zugleich *als man selbst* vorstellt, selber gar nicht vorkommt. Doch in diesem Fall würde man sich gerade nicht vorstellen, dass man selbst eine andere Person ist, sondern einfach eine andere Person – sobald man sich jedoch zusätzlich vorstellt, man selbst sei identisch mit dieser Person, ist man darauf festgelegt, dass man sich selbst als diese Person vorstellt.

Wenn wir uns vorstellen wollen, eine andere Person zu sein, schwanken wir also zwischen zwei Alternativen: Entweder wir stellen uns vor, wie es (in möglicherweise umfassender Art) *für uns* wäre, die Eigenschaften oder Erlebnisse einer anderen Person zu haben. Oder wir stellen uns einfach eine andere Person mit bestimmten Eigenschaften und Erlebnissen vor. Doch damit stellen wir uns dann, wie hoffentlich deutlich geworden ist, nichts mehr über uns selbst vor – und daher auch nicht, dass wir diese Person sind.

Ist das Verstehen von anderen auf eine starke Lesart angewiesen?

Im Gegensatz zu der starken Lesart (man stellt sich vor, eine buchstäblich andere Person zu sein) ist es durchaus möglich, sich vorzustellen, man sei (in umfassender Weise) *wie* eine andere Person (schwache Lesart). In gewisser Weise ist dies ein unbefriedigendes Ergebnis, da so viele Versuche, andere Personen – seien sie real oder fiktiv – zu verstehen, darauf angewiesen zu sein scheinen, dass man sich genau das vorstellen kann: nämlich eine *andere* Person zu sein. Der Erkenntniswert, den wir daraus ziehen, dass wir unsere Sicht auf die Welt beiseite schieben und stattdessen die Perspektive einer anderen Person einnehmen, beruht doch darauf, könnte man meinen, dass man gerade nicht darüber nachdenkt, wie die Situation des anderen für einen selbst wäre, sondern wie sie für *ihn* ist. Und auch die ästhetischen Freuden, das Mitleiden und Mitfühlen mit fiktiven Charakteren im Film und in der Literatur, scheinen darauf angewiesen zu sein, dass man sich der Illusion hingibt und die eigene Perspektive (also die eines Zuschauers oder Lesers) hintenanstellt.

Doch für die Rolle, die die Fähigkeit, sich vorzustellen, eine andere Person zu sein, in unserem Verstehen von anderen Menschen und Werken der Kunst

spielt, ist es weder notwendig noch auch nur förderlich, diese Vorstellung als etwas zu begreifen, das darüber hinausgeht, sich vorzustellen, *wie* eine andere Person zu sein. Die wichtigen Erkenntnisse über uns und andere, die wir aus solchen Vorstellungsakten ziehen, sind in konstitutiver Weise darauf angewiesen, dass man sich etwas über sich selbst vorstellt, sodass die schwächere Lesart zugleich die bessere darstellt. Walton weist darauf hin, dass man sich, um zu verstehen, wie es für eine Minderheit ist, diskriminiert zu werden, nicht Fälle von Diskriminierung vorstellen muss, sondern wie man sich fühlen würde oder wie es für einen wäre, selbst diskriminiert zu werden. „It is when I imagine *myself* in another's shoes (…) that my imagination helps me to understand *him*." (Walton 1990, 34) Das Abrücken von der eigenen Perspektive wird dadurch geleistet, dass man sich die Umstände der Situation einer *anderen Person* vorstellt. Damit ich jemand anderen verstehen kann, selbst wenn ich die Situation anders empfunden habe oder hätte, muss ich jedoch Gefühle dieser Art kennen, d.h. wissen wie es ist, wenn man *sich selbst* so fühlt oder etwas auf diese Weise empfindet. Die Perspektivenverschiebung, die das Verstehen einer anderen Person gewährleistet, besteht folglich darin, sich vorzustellen, dass man sich in der Situation der anderen Person befindet; um zu verstehen, *wie* es für die andere Person ist, in dieser Situation zu sein, ist es aber notwendig, dass man sich Gefühle von der Art, wie der andere sie hat, für sich selbst vorstellt. Denn zu der spezifischen Erlebnisperspektive einer Situation hat eben nur die Person, deren Perspektive es ist, einen privilegierten Zugang. Daher bleibt gar keine andere Möglichkeit, als mir vorzustellen, wie es für mich (womöglich unter anderen Vorzeichen) wäre, in der entsprechenden Situation zu stecken. Gerade die Tatsache, dass ich nicht die identische Perspektive, sondern nur eine nachvollziehende Sicht auf die Dinge habe, kann mir die Differenz zwischen mir und dem Anderen bewusst machen. Das Verstehen der Lebenssituation von anderen Menschen zielt jedoch genau darauf: den anderen als Anderen zu verstehen. Der Erkenntnisgewinn, samt seinen möglichen Konsequenzen für unser eigenes Handeln, wird daraus gespeist, dass wir uns im Nachvollzug der Perspektive des Anderen der *Gemeinsamkeit und Unterschiedlichkeit* zu unserer eigenen Perspektive bewusst bleiben. Damit meine Vorstellung davon, wie jemand anderes eine Situation wahrgenommen oder empfunden hat, für mich einen Grund darstellt, ihn in *meinen Überlegungen* oder in *meinem Handeln* zu berücksichtigen, darf mir die Sicht des anderen nicht als meine eigene erscheinen.

Auch der ästhetische Genuss, der aus dem Sich-Überlassen an eine imaginierte Situation entsteht, ist nicht darauf angewiesen, dass es sich nicht um eine (womöglich sogar vollständige) Illusion handelt. Viele Freuden des Nachvollzugs von fiktiven Ereignissen beruhen darauf, dass es sich um eine *Korrespondenz* und nicht um eine *Konvergenz* der Perspektiven der Zuschauer oder Leser und

der Figuren der Fiktion handelt (vgl. Seel 2013, 214–223). Den Schrecken eines Horrorfilms zu genießen, bedeutet sicher nicht, sich vorzustellen, man sei eine der beteiligten Personen; wäre dies so, würde wohl niemand freiwillig in einen solchen Film gehen. Und auch mein suggestives Eingangsbeispiel zeigt bei näherer Betrachtung, dass wir bei der Lektüre von *Madame Bovary* nicht nur, wie Emma, Charles behäbige Gewöhnlichkeit in Gestalt seiner dicken Finger verachten, sondern vielmehr zugleich darüber erschrecken, dass sie so fühlt oder dass uns ähnliche Regungen vielleicht nicht fremd sind. Und dieses Erschrecken, das wesentlicher Teil der ästhetischen Erfahrung ist, ist eben nicht Teil von Emmas Erleben der Situation, sondern entspringt aus dem Verhältnis zwischen dem Mitfühlen-mit-ihr und dem gleichzeitigen Bewusst-bleiben meiner eigenen Empfindungen. Sowohl die ästhetische Erfahrung als auch die Erkenntnisse, die sich aus Fiktionen ziehen lassen, leben von dem Bewusstsein davon, dass die Imagination nur eine Perspektivenverschiebung und kein Perspektivenwechsel ist.

Sich in eine andere Person hinein zu versetzen, bedeutet also, sich vorzustellen, in relevanten Hinsichten wie jemand anderes zu sein oder zu empfinden. Und dies bedeutet, sich vorzustellen, wie es für *einen selbst wäre, an der Stelle einer anderen Person zu sein oder so zu empfinden wie der Andere*, und nicht, *diese andere Person zu sein.* Wer hingegen behauptet, sich (und sei es in einem minimalen Sinn) vorstellen zu können, eine andere Person zu sein, täuscht sich gewaltig nicht nur über seine eigene, sondern über die Vorstellungskraft überhaupt. Und zwar aus demselben Grund wie jeder, der glaubt, sich eine logische Unmöglichkeit vorstellen zu können: „The reason why some can conceive a barber who shaves all and only the non-self-shavers, while others find this inconceivable, is that the first group needs to learn more logic." (Yablo 1993, 39f.)[4]

Literatur

Chalmers, David 2002: "Does Conceivability Entail Possibility?". In: Tamar Gendler & John Hawthorne (eds.), *Conceivability and Possibility*. Oxford: Oxford University Press, 145–200.

McGinn, Colin 2004: *Mindsight. Image, Dream, Meaning*. Cambridge, Mass.: Harvard University Press.

4. Dieser Essay wurde während eines Stipendiums der Studienstiftung des deutschen Volkes angefertigt, bei der ich mich herzlich für die Förderung bedanke. Danken möchte ich zudem den anonymen Gutachtern der Gesellschaft für Analytische Philosophie für hilfreiche Kommentare, sowie Gerson Reuter, Jochen Schuff und Martin Seel für hilfreichen Widerspruch.

Seel, Martin 2013: *Die Künste des Kinos*. Frankfurt/M.: Fischer.
Walton, Kendall 1990: *Mimesis as Make-Believe*. Cambridge, Mass.: Harvard University Press.
Yablo, Stephen 1993: "Is Conceivability a Guide to Possibility?". *Philosophy and Phenomenological Research* 53, 1–42.

Grazer Philosophische Studien
90 (2014), 327–334.

KANN ICH MIR VORSTELLEN, EINE ANDERE PERSON ZU SEIN?

Viktoria KNOLL
Universität Hamburg

Gewinnerin des dritten Preises des Essay-Wettbewerbs für Studierende 2013
gesponsert durch die Gesellschaft für Analytische Philosophie (GAP)
in Kooperation mit den Grazer Philosophischen Studien*

Summary

For almost every other person than me, when told to imagine being identical to her, I cannot do as ordered. In this essay, I will argue that, although it is possible for me to imagine *myself* to be in a situation which I ascribe to this person, and to imagine having some of her properties, this is not sufficient for imagining being *identical* to her. However, whereas it is impossible for me to imagine being identical to a person who is unlike myself in all relevant respects, it does seem possible for me to imagine being identical to a duplicate of mine.

> "I must be Mabel after all, and I shall have to go and live in that poky little house, and have next to no toys to play with, and oh! ever so many lessons to learn!"
>
> (Alice in Lewis Carroll's
> *Alice's Adventures in Wonderland*)

Wer *Alice im Wunderland* kennt (und gerade PhilosophInnen sei das geraten!), der weiß, dass dort von den erstaunlichsten Dingen die Rede ist: Da gibt es ein sprechendes Kaninchen unter Zeitdruck, den traurigen Schildkrötensupperich (halb Kalb, halb Schildkröte), die altbekannte Grinsekatze, deren Grinsen noch da ist, wenn sie schon fort ist, und vieles, vieles mehr. Es wimmelt geradezu

* Die Preisfrage des Wettbewerbs 2013 war: „Kann ich mir vorstellen, eine andere Person zu sein?" Unter den 36 Einreichungen, die den Regeln des Wettbewerbs entsprachen, wählte die Jury drei Beiträge aus, die den jeweils ersten, zweiten und dritten Platz belegten. Die Autorinnen und Autoren der Gewinnerbeiträge erhielten die Erlaubnis, ihre Essays für die endgültige Publikation geringfügig zu überarbeiten.

von Absurditäten. Und mittendrin steckt die kleine Alice, die von alledem in fürchterliche Verwirrung gestürzt wird. Unter anderem ist sie in Carrolls Geschichte z. B. kurzzeitig überzeugt davon, Mabel zu sein (etwas, was ganz und gar schrecklich für sie ist, denn Mabel ist ein unheimlich dummes Kind). Alice stellt sich vor, dieses andere Mädchen zu sein, daher die einfachsten Dinge nachlernen und ab jetzt in einem schäbigen Häuschen leben zu müssen – und bricht in Tränen aus. Welche Vorstellung aber genau rührt sie hier eigentlich zu Tränen? Kann sie sich überhaupt vorstellen, eine andere Person als sie selbst zu sein? Klar ist: Alice' Bekümmerung ist unbegründet, denn sie und Mabel sind zwei verschiedene Personen. Ganz nach Carrolls Geschmack wäre es doch aber, wenn Alice sich nicht nur täuscht, sondern auch ihre beschriebene *Vorstellung* eine ist, die man außerhalb von Geschichten unmöglich haben kann. Wie ist es also? Kann man sich, kann *ich* mir tatsächlich vorstellen, eine andere Person zu sein?

Zur Beantwortung dieser Frage gilt es zunächst festzulegen, wie wir sie verstanden wissen wollen: „Kann" fragt wohl nach einer meiner tatsächlichen Fähigkeiten, „ich" scheint mich *als Person* herauszugreifen (denn ich soll mir ja gerade vorstellen, eine *andere* Person zu sein) und „sein" steht anscheinend kurz für „identisch sein mit". Interessanter ist da schon die Frage, wie das „andere" verstanden werden soll, allerdings ist wohl auch das schnell geklärt. Denn es scheint offensichtlich, dass ich mir mich selbst mit einigen Eigenschaften vorstellen kann, die mir eigentlich nicht zukommen: Ich, in meiner Vorstellung in naturblond und mit einer Körpergröße von 1,80 m – das ist kein Problem. Um die Frage also nicht zu trivialisieren, sollte das „andere" in ihr nicht als „dieselbe (Person), nur mit anderen Eigenschaften" gelesen werden (wie das umgangssprachlich oft passiert; vgl. „Jetzt, nach seiner Heirat, ist er ein ganz anderer Mensch!"), sondern wohl besser als „nicht mit mir selbst identische" verstanden werden. Zwischenstand also: Habe ich (Person VK) die Fähigkeit, mir vorzustellen, dass ich mit einer Person identisch bin, die nicht identisch mit mir selbst ist? An der Fragestellung zu klären, bleibt jetzt noch „vorstellen" und „Person".

Der ausufernden philosophischen Debatte um den Begriff der Person möchte ich mich in diesem Essay jedoch (soweit möglich) entziehen. Denn uns allen scheint übereinstimmend intuitiv klar zu sein, wann ein Exemplar unter den Begriff *Person* fällt, und wann nicht. Wir scheinen in dieser Hinsicht lediglich dann unsicher und geteilter Meinung zu sein, wenn wir uns in Grenzgebiete um Embryos, KomapatientInnen und ähnliche begeben, die für diesen Essay jedoch ausgeblendet werden können. „Person" soll ferner natürlich nicht kurz für „juristische Person" stehen und scheint zudem auch keine verschiedenen sprachlichen Bedeutungen zu haben, so dass zu klären wäre, welche hier einschlägig ist. Hilfreich ist es jedoch sicher, die Essayfrage an einem Beispiel zu untersuchen.

Und so wollen wir uns im Folgenden fragen: Kann ich, die Autorin, mir vorstellen, mit der Person Charles Lutwidge Dodgson alias Lewis Carroll (fortan LC) identisch zu sein? Wie steht es nun schließlich mit dem letzten ungeklärten Teil der Frage, dem Wörtchen „vorstellen"?

Hören wir „sich etwas vorstellen", denken wir vielleicht zunächst nur daran, dass uns vor unserem geistigen Auge ein bestimmter Gegenstand erscheint. Der Gehalt dieser ersten Art von Vorstellungen ist die mentale Repräsentation eines erdachten oder tatsächlichen Dinges, so dass uns beim Vorstellen ein (bewegtes) Bild erscheint, das einen bestimmten Gegenstand abbildet. Aufgefordert, sich LC vorzustellen, erscheint uns so als gegenständliche Vorstellung z. B. das Bild eines bestimmten Mannes. Und aufgefordert, mir das Ereignis „Händeschütteln von LC und VK" vorzustellen, erscheint mir im Geiste das (bewegte) Bild von LC und mir selbst, das beide so beim Händeschütteln zeigt, als ob jemand Drittes dieses Bild aufgenommen hätte.

Diese so gearteten Vorstellungen lassen sich nun von Vorstellungen einer zweiten Art zumindest grob unterscheiden. Vorstellungen dieser zweiten Art sind solche, deren Gehalt das Erleben bestimmter (Sinnes-)Eindrücke (Gefühle, Gerüche, Geschmäcker etc.), das Ausführen einer bestimmten Handlung oder auch das Haben einer bestimmten Überzeugung *aus der Perspektive des/der Vorstellenden* ist. Was man sich in diesem zweiten Fall vorstellt, ist nicht (allein) die gegenständliche Repräsentation eines Dinges, sondern man stellt sich bei dieser Art vor, etwas „von innen", das heißt aus der Perspektive der 1. Person Singular, zu hören, zu fühlen, zu spüren bzw. eine bestimmte Überzeugung zu haben, oder etwas zu tun. Die Aufforderung, mir das oben angesprochene Ereignis des Händeschüttelns vorzustellen, muss in mir nun keine Vorstellung erster Art hervorrufen, deren Gehalt die Repräsentation von LC und mir aus der Außenperspektive ist. Vielmehr kann mit der Aufforderung auch eine Vorstellung dieser zweiten Art beschworen werden, so dass vor meinem inneren Auge ein Film abläuft, der aus meiner Perspektive gedreht ist. Der Gehalt meiner Vorstellung ist das Erleben des Händeschüttelns aus der Perspektive der 1. Person Singular: Ich sehe meine Fußspitzen, meine zum Schütteln ausgestreckte Hand, LC, und wie er meine Hand in die seine nimmt und schüttelt; der Gehalt der Vorstellung kann ferner das Spüren des Händedrucks, die Wahrnehmung seines Parfums oder andere Sinneseindrücke umfassen, die in der Vorstellung aus der Innenperspektive wahrgenommen werden.

Die Frage, ob ich mir vorstellen kann, mit LC identisch zu sein, kann somit auf zweierlei Weise präzisiert werden: Kann ich mir mithilfe einer Vorstellung erster Art vorstellen, dass ich mit LC identisch bin? Und: Kann ich mir auf unsere zweite Art vorstellen, mit ihm identisch zu sein? Beide Fragen gilt es im Folgenden zu klären.

Frage 1 ist dabei jedoch, meines Erachtens, schnell beantwortet. Denn wenn der Gehalt meiner Vorstellung lediglich die mentale Repräsentation meiner Wenigkeit und/oder der LCs aus einer dritten Perspektive ist, so gibt es in meiner Vorstellung keinerlei relevante Verbindung zwischen mir und ihm, die es rechtfertigen würde, hier von der Vorstellung zu sprechen, dass ich mit ihm identisch bin. Ich stelle mir dann bloß mich selbst und/oder LC aus der Außenperspektive vor (vgl. auch Williams 1966, 43). Das heißt, wollen wir „vorstellen" auf diese erste Art verstehen, ist unsere Essayfrage schnell beantwortet, und zwar mit „Nein": Ich habe nicht die Fähigkeit, mir mithilfe einer gegenständlichen Vorstellung erster Art vorzustellen, dass ich mit LC identisch bin. Denn hier stelle ich mir lediglich mich selbst, LC oder uns beide zusammen vor.

Bleibt die zweite genannte Art des Vorstellens, ein Vorstellen des Erlebens von Eindrücken, des Habens einer Überzeugung oder des Ausführens von Handlungen aus der Perspektive der 1. Person Singular. Was würde diese Lesart unserer Frage an Vorstellungsleistung von mir verlangen? Sie würde wohl verlangen, mir Eindrücke und Überzeugungen, wie LC sie gehabt hat, *als LC* vorzustellen, bzw. im Geiste als LC Handlungen auszuführen. Und tatsächlich ist es diese Art des Vorstellens, die in der Literatur als einschlägig für die Vorstellung angesehen wird, mit einer anderen Person identisch zu sein. So beschreibt Reynolds z. B., wie er sich vorstellt, Napoleon auf dem Schlachtfeld zu sein (Reynolds 1989, 616): "I ‚see' the horse on which I am mounted and the armies clashing in the distance, I ‚hear' the hoofbeats […]; I ‚smell' the smoke of gun powder" – und setzt gleich von Anfang an voraus, dass es selbstredend möglich ist, sich auf diese Weise vorzustellen, eine andere Person zu sein, genau wie viele andere dies in dieser Debatte tun (z. B. Williams 1966, Nichols 2008, Ninan 2009). Und es scheint natürlich auch einleuchtend, dass ich mir mithilfe einer Vorstellung zweiter Art vorstellen kann, Eindrücke oder Überzeugungen zu haben bzw. Handlungen auszuführen, die ich LC zuschreibe. Ich kann mir mit einiger Mühe durchaus vorstellen, aus meiner Perspektive eine selbstgemachte Zeichnung von Alice vor mir auf dem Papier zu sehen, eine Hand in die Themse zu halten (und das Wasser zu spüren) und religiöse Überzeugungen zu haben. Es ist jedoch fraglich, ob das tatsächlich schon ausreicht, um davon sprechen zu können, dass ich mir hier vorstelle, identisch zu sein mit LC. Mir scheint es nämlich plausibler anzunehmen, dass ich mir in einem solchen Fall nicht vorstelle, *identisch* zu sein mit LC, sondern dass ich mir hier vielmehr vorstelle, *ich* sei in einer Situation, von der ich stipuliere, dass sie auch LC durchlebt haben muss. Die Vorstellungsleistung, aus der Innenperspektive stipulierte Erlebnisse von Carroll zu durchleben bzw. Überzeugungen zu haben, die auch er hatte, scheint mir nicht die hier behandelte Essayfrage positiv zu beantworten, sondern vielmehr zu bestätigen, dass

es möglich ist, mir auf zweite Art vorzustellen, *ich selbst* sei in der Situation des Schriftstellers, stecke in seiner Haut und die Hand in die Themse.

Die Frage, wie eine solche Vorstellung zweiter Art von mir selbst in der Situation einer anderen Person handeln kann, obwohl der Gegenstand meiner Vorstellung (der, aus dessen Perspektive in meiner Vorstellung etwas erlebt wird) Eigenschaften hat, die ich tatsächlich nicht habe, ist dabei leicht zu beantworten. Denn es ist zwar richtig, dass ich z. B. keine religiösen Überzeugungen habe, die Person in meiner Vorstellung aber schon. Das heißt aber nicht, dass die Person, aus deren Perspektive in der Vorstellung etwas erlebt wird, nicht ich selbst sein kann. Denn, wie anfangs bereits festgestellt, ist es durchaus möglich, mir vorzustellen, *ich* sei naturblond und 1,80 m groß. Warum sollte es dann nicht auch möglich sein, mir vorzustellen, ich durchlebe die stipulierte Situation von LC *und* sei religiös? Die Eigenschaft, nicht religiös zu sein, ist schließlich auch wohl keine, die mir essentiell, also notwendigerweise zukommt (anders wohl als die Eigenschaft, diese bestimmte Person VK zu sein). Und ferner scheint auch das Durchleben bestimmter Situationen mir nicht essentiell zu sein. Insofern sollte ich mir mich selbst erst recht problemlos ohne diese Eigenschaften vorstellen können. Somit kann also aus der Tatsache, dass die Person in meiner Vorstellung, von deren Innenperspektive meine Vorstellung handelt, andere Eigenschaften hat, als ich selbst tatsächlich habe, nicht gefolgert werden, dass es sich bei dieser Person nicht um mich selbst handelt (in der Situation, wie sie LC meines Erachtens erlebt haben könnte). Und somit folgt ebenso wenig, dass die Essayfrage positiv zu beantworten ist.[1]

Mir auf die hier besprochene zweite Art vorzustellen, ich sei *identisch* mit LC, müsste, meines Erachtens, auch weit über diese gerade beschriebene Vorstellungsleistung hinausgehen. Es reicht nicht aus, mir nur vorzustellen, z. B. einige Überzeugungen LCs bewusst in meiner Vorstellung zu übernehmen; denn LC und mich unterscheidet schließlich mehr als ein paar Überzeugungen. Was ich in meiner Vorstellung also im Hinterkopf haben müsste, ist, dass ich ausreichend viele *seiner* (essentiellen) Eigenschaften in meiner Vorstellung selbst haben muss, und ausreichend viele *meiner* (essentiellen) Eigenschaften in meiner Vorstellung abzulegen habe, damit ich wirklich davon sprechen könnte, in meiner Vorstellung mit LC *identisch* zu sein. Natürlich müsste ich mir (während der Vorstellung z. B. des Habens einiger seiner Überzeugungen aus meiner Perspektive) nicht all dieser Eigenschaften, die ich dann neu hätte, bewusst sein. Aber mir müsste

1. Die Bürde, nachzuweisen, dass ich mir hier nicht mich selbst vorstelle, scheint mir nun auch bei meinen KontrahentInnen zu liegen. Diese müssten somit zeigen, warum die von ihnen *angeblich* beschriebene Vorstellung (ich stelle mir vor, mit einer anderen Person identisch zu sein) nicht eigentlich korrekt so zu beschreiben ist, wie ich es vorschlage (ich stelle mir mich selbst in der Situation einer anderen Person und mit einigen ihrer Eigenschaften vor).

beim Vorstellen bewusst sein, dass nicht mehr ausreichend viele meiner ursprünglichen, eigenen (essentiellen) Eigenschaften in meiner Vorstellung zweiter Art übrig sein dürften, um mit Recht sagen zu dürfen, dass ich mir vorstelle, mit LC identisch zu sein.

Bei solch einer Vorstellung wäre nun aber bereits jegliche Verbindung zu meiner Person abgebrochen – denn genug Eigenes von mir dürfte ja auch in meiner Vorstellung nicht mehr übrig sein! Wie könnte ich so überhaupt noch der Überzeugung sein, dass es sich nun beim Gegenstand meiner Vorstellung noch um mich selbst, als den Gegenstand, der identisch mit LC sein soll, handelt? Dass der Gegenstand, aus dessen Perspektive in meiner Vorstellung erlebt wird, also tatsächlich noch ich wäre? Was ich mir hier doch eigentlich am Ende versuchen würde vorzustellen, wäre schlicht das Erleben der möglichst vollständig vorgestellten Perspektive LCs aus der Perspektive der 1. Person Singular – aber eben nicht mehr, mit ihm identisch zu sein. Denn damit Letzteres wirklich erfüllt sein könnte, müsste in meiner Vorstellung eine Relation zwischen zwei Gegenständen hergestellt werden (die Relation der Identität zwischen VK und LC), wozu es in der Vorstellung des, irgendwie gearteten, Vorhandenseins beider Relata bedürfte.

Auch die zweite Präzisierung unserer Essayfrage kann damit also mit „Nein" beantwortet werden: Ich kann mir auch mithilfe einer Vorstellung zweiter Art nicht vorstellen, dass ich mit LC identisch bin. Denn entweder stelle ich mir dann, aus der Perspektive der 1. Person Singular, bloß mich selbst in der Situation von LC vor (mit einigen seiner Eigenschaften ausgestattet) oder von mir selbst ist in meiner Vorstellung aus der Innenperspektive nicht mehr genug übrig, um mit Recht von der Vorstellung sprechen zu dürfen, *ich* sei mit LC identisch; denn eines der Relata, zwischen denen sich die Relation der Identität vorgestellt werden soll, wäre unterrepräsentiert.

Eine Erwiderung auf diese Argumentation liegt nun aber auf der Hand: Selbst wenn überzeugend gezeigt wurde, dass ich mir auf keine der beiden Arten vorstellen kann, mit *LC* identisch zu sein, heißt das nicht, dass die Essayfrage prinzipiell mit „Nein" zu beantworten ist. Es kommt schlicht darauf an, an welchem Beispiel die Frage untersucht wird! Es mag überzeugend sein, dass ich mir nicht vorstellen kann, mit einer Person identisch zu sein, die mir selbst sehr unähnlich ist. Wie ist es aber, wenn die Person, mit der ich in meiner Vorstellung identisch sein soll, zwar numerisch verschieden von mir ist, jedoch maximale qualitative Ähnlichkeit mit mir aufweist? Nehmen wir z. B. an, ich steige in Parfits Teletransportationsapparat (vgl. Parfit 1995), es geht etwas schief und ich werde deshalb nicht wie geplant zum Mars teletransportiert, sondern es wird lediglich ein Duplikat D von mir erstellt, das auf einmal direkt neben mir im Apparat erscheint. D soll mir dabei nicht nur äußerlich bis auf's Haar gleichen,

sondern mir auch in seiner mentalen Ausstattung maximal ähnlich sein. Was stelle ich mir nun vor, wenn ich aufgefordert werde, mir vorzustellen, ich sei mit D identisch?

Würde es sich bei D um eine Person handeln, die mir wie LC höchst unähnlich ist, so könnte, wie oben bemerkt, nicht mit Recht davon gesprochen werden, dass ich mir auch wirklich das vorstelle, wozu ich aufgefordert wurde. Nicht nur Ds Situation aber wurde hier ja nun gerade so konzipiert, dass sie dieselbe wie meine ist; vor allem auch D zu sein, scheint sich aus der Perspektive der 1. Person Singular nicht davon zu unterscheiden, VK zu sein. Wieso sollte es auch? Körperlich und mental besteht im Gedankenexperiment schließlich maximale Ähnlichkeit zwischen D und mir. Stelle ich mir nun mich selbst in Ds (bzw. meiner) Situation vor, dann ist anscheinend der Gehalt meiner Vorstellung derselbe wie der Gehalt von Ds Vorstellung von sich in seiner (bzw. meiner) Situation. Ferner scheint der Gehalt von Ds Vorstellung von sich in jeder beliebigen Situation aufgrund unserer enormen Ähnlichkeit derselbe zu sein wie der Gehalt meiner Vorstellung von mir in derselben Situation.[2]

Bei LC und mir dagegen ist diese Gehaltsgleichheit der Vorstellungen nicht gegeben: Der Gehalt meiner Vorstellung von mir in irgendeiner Situation ist *nicht* derselbe wie der Gehalt von LCs Vorstellung von sich in derselben Situation. Denn, wie schon bemerkt, unterscheiden sich LC und ich stark voneinander, so dass die Vorstellung von sich selbst je eine völlig andere ist. Versuche ich mir, mir dieser Unähnlichkeit bewusst, vorzustellen, dass ich mit LC *identisch* bin, dann kann ich deshalb nur scheitern. Denn in meiner Vorstellung ist von mir selbst dann schlicht nicht mehr genug übrig.

Im Gedankenexperiment aber liegt eine andere Situation vor. Denn der Gehalt meiner Vorstellung von mir in einer Situation scheint derselbe zu sein wie der Gehalt von Ds Vorstellung von sich in derselben Situation. Denn D zu sein ist in der Vorstellung genauso, wie VK zu sein. Und somit sind beide Personen, D und VK, in meiner Vorstellung von mir selbst in Ds Situation repräsentiert. Beide Relata, zwischen denen die Relation der Identität in der Vorstellung hergestellt werden soll, sind so Teil meiner Vorstellung. Und das kann nun durchaus als überzeugendes Indiz dafür gewertet werden, dass meine Vorstellung von mir in Ds Situation tatsächlich auch mit Recht als „Vorstellung mit D identisch zu sein" bezeichnet werden kann – genau wie auch Ds Vorstellung von sich in derselben

2. Ein anderes Beispiel, in dem verschiedene Vorstellungsvorkommnisse zweiter Art denselben Gehalt haben, ist das Beispiel der Vorstellung, *Highway 61 Revisited* mit geschlossenen Augen, abgespielt von CD-Exemplar 1 zu hören und der Vorstellung, dasselbe Album mit geschlossenen Augen, abgespielt von CD-Exemplar 2 zu hören. Was man sich hier bei beiden Vorstellungen zweiter Art jeweils vorstellt, scheint identisch zu sein (sofern beide CD-Exemplare tadellos abspielen), derselbe Gehalt kann jedoch auf unterschiedliche Weise beschrieben werden.

Situation mit Recht den Namen „Vorstellung mit VK identisch zu sein“ tragen darf. Beide Vorstellungen haben schlicht denselben Gehalt.

Wenn wir hier deshalb glauben, dass D tatsächlich eine andere Person als VK ist (und das scheint aufgrund der Tatsache, dass ich und D zumindest numerisch verschieden sind, erst einmal vernünftig zu sein), dann mögen wir, angesichts dieses Gedankenexperiments, einen Fall für eine andere Person gefunden haben, von der ich mir vorstellen kann, mit ihr identisch zu sein.

Lewis Carroll aber führt uns natürlich als LeserInnen letztlich trotzdem in die Irre und bleibt damit der Meister des Absurden. Denn, wie dieser Essay zeigen sollte, schreibt er Alice schließlich eine Vorstellung zu, die das Mädchen so unmöglich haben kann. Denn sie stellt sich in der Geschichte ja gerade vor, mit *Mabel* (und eben nicht mit einem Duplikat von sich selbst) identisch zu sein und gerät darüber in Verzweiflung. Carrolls absurdem Charme an der zitierten Stelle wird man sich, meines Erachtens, nur nicht gleich gewahr (anders als im Fall von Grinsekatze & Co.), weil wir das von Carroll wortwörtlich Niedergeschriebene leicht in etwas übersetzen können, was Alice sich (auch als nicht-fiktionale Person) ohne Weiteres, mithilfe einer Vorstellung zweiter Art, vorstellen könnte: Nämlich sich *selbst,* Alice, in der Situation der schrecklich dummen Mabel.

Literatur

Nichols, Shaun 2008: “Imagination and the *I*”. *Mind & Language* 23, 518–535.

Ninan, Dilip 2009: “Persistence and the First-Person Perspective”. *Philosophical Review* 118, 425–464.

Parfit, Derek A. 1995: “The Unimportance of Identity”. In: Henry Harris (Hg.), *Identity*. Oxford: Oxford University Press, 13–45.

Reynolds, Steven L. 1989: “Imagining Oneself to Be Another”. *Nous* 23, 615–633.

Williams, Bernard 1966: “Imagination and the Self”. In: Bernard Williams (Hg.), *Problems of the Self – Philosophical Papers 1956–1972.* Cambridge: Cambridge University Press, 26–45.

Grazer Philosophische Studien
90 (2014), 335–346.

IS EPISTEMOLOGICAL DISJUNCTIVISM THE HOLY GRAIL?[1]

Guido MELCHIOR
University of Graz

In his highly valuable book, Duncan Pritchard presents a particular account of perceptual knowledge, *epistemological disjunctivism* (ED). Pritchard argues that this view seems plainly false at first sight, but if it were right, it would represent the "holy grail of epistemology" (1), a view that allows us "to have our cake and eat it too" (3). This prospect motivates Pritchard to develop and defend an account that prima facie might seem simply false. It is disputable whether ED really seems plainly false at first sight or whether this intuition is based on a particular philosophical tradition. However, in this review I will not discuss whether ED is actually true. Rather, I will investigate whether, if true, it has the advantages over rival accounts that Pritchard claims.

One attributed advantage of ED is that it has the potential to dissolve the conflict between epistemic internalism and epistemic externalism, which Pritchard characterizes as follows: epistemic internalism is the view that the crucial epistemic factors for knowledge and justification are internal to agents and, therefore, reflectively accessible to them. According to epistemic externalism, crucial epistemic factors are external and world-linked. Internalism has the advantage of being able to explain the concept of epistemic responsibility, but faces the problem that the epistemic standing of beliefs is not truth-linked. In contrast, externalism can easily establish a connection between the epistemic standing of beliefs and their truth, but faces serious problems in explaining epistemic responsibility. Pritchard claims that ED can overcome this impasse by adopting elements of internalism *and* externalism. He defines ED as follows:

> Epistemological Disjunctivism: The Core Thesis
> In paradigmatic cases of perceptual knowledge an agent, S, has perceptual knowledge that ϕ in virtue of being in possession of rational support, R, for her belief that ϕ which is both *factive* (i.e., R's obtaining entails ϕ) and *reflectively accessible* to S. (13)

---

1. Review essay of Duncan Pritchard, *Epistemological Disjunctivism*. Oxford: Oxford University Press. 2012. 206 pp. ISBN 9780199557912.

Thus, ED combines the externalist element of factive support with the internalist element of reflective accessibility. Crucially, ED rejects the standard claim about perceptual knowledge that agents have the same degree of *reflectively accessible rational support* in good cases of veridical perception and in bad cases of illusions or hallucinations. Rather, ED claims that this support is radically different in the two cases (15).

ED is a view about *perceptual* knowledge, not about knowledge in general. In paradigm cases of perceptual knowledge, the particular factive rational support that is also reflectively accessible is our "*seeing* that the target proposition obtains" (14). Pritchard points out that ED is in accordance with "commonsense ways of thinking, and talking, about perceptual reasons" (17) according to which it is normal to say that I know that p because I see that p. He argues that any philosophical view that is in accordance with commonsense enjoys a default status that is denied to revisionary philosophical views that diverge from common sense. So if ED turns out not to be a complete non-starter as a theoretical position, then it has strong methodological advantages over its revisionary alternatives. Accordingly, Pritchard's main strategy is moderate. He does not want to argue directly for the truth of ED. Rather his objective is to set out what ED amounts to and to explain that ED does not face those prima facie problems that it seems to face. Thereby, Pritchard aims at establishing ED as a possible view that has to be taken seriously into account, especially given its overall attractiveness.

Obviously, the relation between seeing and knowing is crucial for ED. However, Pritchard argues that seeing that p is neither a particular way of knowing that p nor is it sufficient for knowing that p (25). One explanation is that seeing that p does not entail believing that p, whereas knowing that p does. However, seeing that p still guarantees that one is in a good position for knowing that p; a correct analysis, I think.

Pritchard presents a taxonomy of six different cases for exploring the relation between seeing and perceptual knowledge in more detail (29). *Good+* is the case of truly believing that p based on seeing that p in the absence of any defeater. In this case, the epistemic conditions are objectively and subjectively good. *Good+* is a case of veridical perception for Pritchard, where S sees that p and knows that p. In the following five cases, S does not *know* that p anymore. In the second case, *Good*, S sees that p but is or should be in possession of a defeater for p. Here, the epistemic conditions are still objectively good, but *subjectively bad*. In case *Good*, S still sees that p, but does not know that p. In the third case, *Bad*, the circumstances of perception are *objectively bad*, but subjectively good. Such a case occurs, for example, if one is confronted with a real barn in fake-barn county. In the fourth case *Bad+*, the epistemic conditions are neither objectively nor subjectively good. In all four of these cases the perception is veridical, but

the epistemic circumstances vary. In the last two cases, *Bad++* and *Bad+++,* the perception is *not* veridical. In case *Bad++* the epistemic conditions are at least subjectively good, whereas in *Bad+++* S also is or should be aware of a defeater.

Epistemologists often only distinguish between good cases of veridical perception and bad cases, without developing a more fine grained taxonomy. In this respect, Pritchard's taxonomy has its merits. However, I do not find it entirely plausible. In case *Good*, S sees that p, but does not know that p, because S is or should be aware of a defeater, for example, if S is not in fake barn county but a normally trustworthy person told her that she is. In case *Bad,* S actually is in bad epistemic conditions like in fake barn county. In both *Good* and *Bad,* S's perception is veridical, i.e. she is confronted with a barn, but only in case *Good* does S see that there is barn. I find this distinction awkward. It is more plausible that S sees in both cases that there is a barn. Whether S actually is in barn county only affects S's *knowledge* or *justification*. However, not much seems to hinge on this taxonomy. Pritchard could easily adapt it without weakening his point that seeing is not sufficient for knowing, though he would have to weaken his claim about the epistemic position one is in when seeing that p. If S can also see a barn in fake barn county, then S's seeing that p can fail to constitute knowledge for these objective reasons as well. However, this seems like a terminological discussion.

Pritchard's *Epistemological Disjunctivism* consists of three parts. Part one sets out the position of ED and outlines three main prima facie problems that conflict with the view: the *access problem*, the *basis problem* and the *distinguishability problem*. The access problem takes the form of a reductio argument against ED according to which ED entails the implausible claim that one can have knowledge of empirical propositions by a priori reflection alone. Here is how this problem arises according to Pritchard. Suppose that the rational support for my belief that Ann is in the office is that I see her in the office, which is, according to ED, reflectively accessible. Given that I know that seeing that p entails p, it seems that I can deductively conclude from a priori reflection alone that Ann is in her office.

I do not find the access problem convincing, since it seems obvious that, according to ED, seeing that p as a premise of the access problem is more than what there is accessible by reflection alone. However, Pritchard argues sophisticatedly against the access problem. He concludes that the only point it can make is that in case of veridical perception, S can know by reflection alone that her reason for believing that p is the factive empirical reason R (her seeing that p) which entails p. However, this conclusion is far weaker than the conclusion of the original access problem. Pritchard's argumentation is careful and precise, but given the unconvincing nature of the access problem, this argumentation itself will not convince many critics that ED is true.

The second prima facie problem for ED that Pritchard diagnoses is the *basis problem*. This problem arises from the question of how seeing can provide the basis of knowing, if seeing *is* a form of knowing. Pritchard's response to this problem is based on his view about the relation between seeing and knowing. Seeing is not a particular kind of knowledge, and, therefore, the basis problem is neutralized. Again, I see nothing wrong with Pritchard's response, but I also do not see the impact of the basis problem for ED.

The third problem is the distinguishability problem. ED accepts that veridical perception and hallucinations are by assumption introspectively indiscriminable. The problem arising for ED is how factive rational support can be reflectively accessible, given that they are introspectively indiscriminable. Part two of *Epistemological Disjunctivism* is almost entirely devoted to the distinguishability problem. It is the hardest problem for ED according to Pritchard. I agree, moreover I think it is intuitively the most appealing one.

Pritchard argues that when it comes to distinguishability, we have to differentiate between various principles. The following two principles are the most interesting ones:

> The Discrimination Principle
> If S has perceptual knowledge that ϕ, and S knows that another (known to be inconsistent) alternative ψ does not obtain, then S must be able to perceptually discriminate between the object at issue in ϕ and the object at issue in ψ. (73)

> The Favouring Principle
> If S (i) knows that ϕ, and (ii) knows that ψ, and (iii) knows that ϕ entails ψ, then S has better evidence in support of her belief that ϕ than for believing that not-ψ. (76)

Pritchard correctly argues that the discrimination principle puts stronger requirements on knowledge than the favouring principle. The illustrative example he uses is Dretske's zebra case. S can have favoring epistemic support for her belief that the animal in the pen is a zebra and not a painted mule (e.g. by having background knowledge about the trustworthiness of the zoo administration) without having the capacity to perceptually discriminate between zebras and painted mules. In this case, S fulfills the favouring principle but not the discrimination principle. Pritchard argues that the discrimination principle can be dismissed as too strong. The reason for our mistaken intuition in the zebra case is the false assumption that the missing evidence is *discriminatory* evidence, so Pritchard argues. Analogously, ED is not the claim that S can *introspectively*

distinguish between case *Good+* and the bad cases. Rather it is the weaker claim that these cases are *reflectively* distinguishable. Importantly, Pritchard argues that this view is not based on ED and should be acceptable to defenders of alternative views. Pritchard analysis of these cases is convincing, but not new. Vogel's (1990) influential work that comes to a similar conclusion could be cited at that point.

Part three of *Epistemological Disjunctivism* is entirely devoted to the skeptical problem. Pritchard points out that there is an essential difference between the zebra case and skepticism. In the zebra case we can know that the animal in the pen is a zebra and not a painted mule if we have additional and independent rational support. We need not have the capacity to perceptually discriminate between zebras and painted mules. However, the skeptic calls everything into question. Therefore, we cannot appeal to this kind of independent background knowledge for ruling out the skeptical alternative of global hallucination according to Pritchard.

One way ED could respond to skepticism is 'simple-minded epistemological disjunctivist Neo-Mooreanism' which argues along the following lines: In case *Good+*, S knows that p because S sees that p. Given that S's rational support for p entails that p, and that S knows that p entails that S is not a brain in vat deceived in falsely believing that p (~BIV), S can conclude by deduction that ~BIV and thereby reflectively distinguish p from BIV. The way I understand this inference is:

Simple-minded ED
I see that p
Therefore, I know that p
I know that p entails ~BIV
Therefore, I know that ~BIV

Pritchard dismisses simple-minded ED by incorporating a defeater condition. He argues that the orthodox view mistakenly treats all error possibilities equally, no matter whether they are epistemically motivated or just raised (125). It is disputable whether this really is the orthodox view about error possibilities. Pritchard objects that we only need to appeal to additional evidence if the error possibility is motivated. If one buys into Pritchard's distinction, then the solution to the skeptical problem is easy at hand. Radical skeptical hypotheses are just raised but not epistemically motivated (126). Thus, one need not appeal to independent rational support to reject them. Pritchard's anti-skeptical strategy can be summarized as follows:

Pritchard's ED
I see that p
If there are no epistemically motivated error possibilities, then if I see that p, then I know that p
There are no epistemically motivated error possibilities
Therefore, I know that p
p entails ~BIV
Therefore, I know that ~BIV

Pritchard's dismissal of the simple-minded ED seems puzzling and unmotivated. If one accepts that the factive rational support is reflectively accessible as ED does, then why shouldn't one utilize it the way simple-minded ED suggests? Admittedly, it is a familiar externalist move to incorporate a defeater condition into the account of knowledge or justification as Goldman's (1979) original formulation of process reliabilism also suggests, but this connection is neither motivated by process reliabilism nor by ED itself.

Notably, Prichard's anti-skeptical strategy mirrors the internalistic strategy of dogmatism as defended by Huemer (2000) and Pryor (2000 and 2004). They argue that, in absence of any defeaters, one's experience as of p gives one prima facie justification that p. This dogmatist strategy takes the following form:

Dogmatism
I have an experience as of p
In absence of any defeaters, if I have an experience as of p, then I am justified to believe that p
There are no defeaters
Therefore, I am justified to believe that p
I know that p entails ~BIV
Therefore, I am justified to believe that ~BIV

Because of Pritchard's treatment of defeaters, he adopts the dogmatist strategy, except he replaces the internalistic concepts of "experience" and "justification" with the factive concepts of "seeing" and "knowing". One might argue that the factivity of the rational support in the case of ED makes the difference to dogmatism. However, this potential difference does not affect the dialectical potential of the two accounts.

After presenting his anti-skeptical response, Pritchard investigates the dialectical situation concerning skepticism. Pritchard distinguishes between *overriding* and *undercutting* anti-skeptical strategies. Overriding anti-skeptical strategies concede that the skeptical intuition is plausible but claim that there are independent

theoretical grounds for rejecting it, since intuitions are only defeasible guides to truth (132). Undercutting anti-skeptical strategies, in contrast, show that, properly understood, the skeptical claims are not intuitive at all. When it comes to analyzing the dialectic between the skeptic and the anti-skeptic, undercutting strategies are preferable to overriding ones. Moreover, Pritchard claims that ED is clearly undercutting. He admits that there might exist various intermediate cases between overriding and undercutting skeptical strategies, but he assumes that the distinction is clear enough for his purposes (133). I doubt that this is the case.

Let's compare ED to internalistic and externalistic alternatives with respect to (1) knowing that p, (2) knowing that ~BIV, and (3) higher-order knowledge that one knows that p. I pick dogmatism as an internalistic account and process reliabilism as an externalistic one, although other externalistic accounts of knowledge like safety or certain virtue epistemological accounts will deliver similar results. Pritchard argues that the skeptical intuition is based on internalistic intuitions. Therefore, externalistic anti-skeptical strategies are by their nature overriding because they reject internalism for independent theoretical reasons (133). Thus, we should at least find a clear criterion for distinguishing the dialectical features of ED from those of process reliabilism. Let's see, first, how these three accounts can explain the possibility of perceptual knowledge:

> Perceptual knowledge for ED
> S has perceptual knowledge that p iff S sees that p and there are no epistemically motivated defeaters for p.

> Perceptual knowledge for dogmatism
> S has perceptual knowledge that p iff S has an experience as of p and believes that p and there do not exist any defeaters for p (and if the further necessary conditions for converting justified beliefs into knowledge are fulfilled).

> Perceptual knowledge for process reliabilism
> S has perceptual knowledge that p iff S truly believes that p based on a reliable process of perception (and if there do not exist any defeaters for p and if the further necessary conditions for converting justified true beliefs into knowledge are fulfilled).

Obviously each of these three accounts of knowledge offers an explanation of how perceptual knowledge is possible that is crucially based on a more general account of knowledge.

How can we know according to these accounts that ~BIV? One way is simply by deduction as follows:

p
Therefore, I am not a BIV deceived in falsely believing that p

This entailment relation holds if the anti-skeptical proposition is understood as ~((I am a BIV that believes that p) & ~p). The three compared knowledge accounts all seem to license this kind of deductive knowledge.

One might argue with Klein (2004) that this conclusion is just equivalent to the disjunction ~(I am a BIV that believes that p) ∨ p which does not tell us anything informative about the conditions of our perception. So one might search for a stronger anti-skeptical thesis like "I correctly experience and believe that p" with the underlying formal structure E(p) & B(p) & p. According to dogmatism, S can draw the following deductive inference:

Knowledge that ~BIV for dogmatism
p (by having an experience as of p)
I have an experience as of p and I believe that p (by introspection)
Therefore, I correctly experience and believe that p (by deduction)
Therefore, I am not a BIV hallucinating that p (by deduction)

Process reliabilists can use the same inference, except that the belief in p must result from a reliable belief forming process. For ED, the inference takes simply the following form:

Knowledge that ~BIV for ED
I see that p
Therefore, I am not a BIV hallucinating that p

Again, all three accounts of knowledge license knowledge in the stronger denial of the skeptical hypotheses via inference from p itself. How can S acquire a more demanding *higher-order* knowledge that she has perceptual knowledge? In this case, S has to know that the conditions for knowledge that these accounts propose are fulfilled and that these conditions are sufficient for knowledge. The inferences take the following line:

Higher-order knowledge for ED
(1) I see that p (by reflection)
(2) There are no epistemically motivated defeaters for p (by reflection)
(3) If (1) and (2) are true, then I have perceptual knowledge that p (by philosophical argumentation)
(4) Therefore, I have perceptual knowledge that p (by deduction)

Higher-order knowledge for dogmatism

(1) I have an experience as of p (by introspection)
(2) There are no defeaters for p (by reflection)
(3) If (1) and (2) are true (and further conditions for transforming justified beliefs into knowledge are fulfilled), then I have perceptual knowledge that p (by philosophical argumentation)
(4) Therefore, I have perceptual knowledge that p (by deduction)

So far, I do not see any crucial difference between the dialectical capacities of ED and of dogmatism as ED's purely internalistic counterpart. Pritchard argues that ED is dialectically superior to other accounts because of its undercutting nature. Notably Pryor (2004, 362) thinks that Moorean arguments based on dogmatism are "persuasively crippled", but that their justificatory structure is flawless. Thus, there is disagreement about the dialectical capacities of very similar accounts. I doubt that ED's dialectical position is strong. Specifically, I do not see why ED is more theoretically attractive than dogmatism.

For process-reliabilism, higher-order knowledge is achieved slightly differently. Vogel (2000) has pointed out that if process reliabilism is true, then one can acquire knowledge about the reliability of a source (and therefore higher-level knowledge) via bootstrapping, i.e., by basing it on knowledge delivered by this source. In the case of knowledge about the reliability of one's own perception, bootstrapping takes the following form:

Higher-order knowledge for process reliabilism

(1) p (reliable process via perception)
(2) I believe that p based on perceiving that p (reliable process via introspection)
(3) My perception that p and my belief that p are correct (logical inference)
(4) Repeat
(5) My perception is reliable (induction)
(6) If my perception is reliable, then my true beliefs based on perception constitute knowledge (by philosophical argumentation)
(7) Therefore, I have perceptual knowledge that p (by deduction)

Bootstrapping as a process of acquiring higher-order knowledge is usually regarded as a problem for externalist accounts, not as an advantage. Vogel argues that process reliabilism sanctions every step of bootstrapping. However, since bootstrapping is an obviously flawed reasoning process, according to Vogel, process reliabilism is false. Thus, Vogel uses bootstrapping as a reductio argument against process reliabilism. Cohen (2002) argues more generally that any account of

knowledge that allows one to have knowledge via a source without having prior knowledge about the reliability of the source suffers from the "easy knowledge problem". This problem takes the following form for a dogmatist response to the skeptic.

(1) I have an experience as of p
(2) Therefore, p
(3) Therefore, I am not a BIV having a hallucination as of p

Thus, process reliabilism and dogmatism both face the problem of easy knowledge in one way or another. One might suspect that this gives ED a crucial advantage over its rival accounts. However, I doubt that this is the case. Take the following inference from ED:

(1) I see that p
(2) Therefore, I am not a BIV having a hallucination as of p

I doubt that those who criticize process reliabilism and dogmatism for facing the easy knowledge problem find this inference a more convincing anti-skeptical move.

To sum up: ED, dogmatism and process reliabilism provide an explanation for how we can have perceptual knowledge. They also allow for knowledge that ~BIV via deduction from perceptual knowledge. The ways of acquiring higher-order knowledge about perceptual knowledge are different but they are all based on perceptual knowledge. However, ED can provide a far more natural explanation for higher-order knowledge than dogmatism or process reliabilism. I think this is its crucial advantage. Nevertheless, the three strategies seem dialectically equally overriding or undercutting. Given this diagnosis, I do not see the respect in which externalist anti-skeptical strategies should be overriding in nature whereas ED is undercutting.

The way we can acquire knowledge that ~BIV and higher-order knowledge based on perceptual knowledge is somehow unsatisfactory. I think the underlying explanation is that the skeptic not only claims that we neither know ~BIV nor have perceptual knowledge. Rather, she suggests that we do not have perceptual knowledge *because* we do not know that ~BIV. The skeptic thereby implicitly assumes that we need to have knowledge that ~BIV in the first place in order to have any kind of perceptual knowledge. This view is *conservatism,* as opposed to *liberalism*, about perception. Conservative anti-skeptics like Vogel (1990a) think that we have the kind of necessary a priori knowledge that ~BIV. Importantly, the skeptic is also a conservative but she rejects the view that we have any a priori

justification for ~BIV. By making its conservative presupposition explicit, we can formulate skepticism as follows:

> Conservative skepticism
> We need to have prior knowledge that ~BIV in order to have perceptual knowledge that p.

Liberalism about perception allows for basic perceptual knowledge to take different forms. Process reliabilism licenses basic knowledge from any reliable source. Pryor, as a dogmatist, is a liberal about perception but thinks that other sources should be treated differently. ED is a view only about perceptual knowledge and does not make any claims about other potential knowledge sources. The three accounts reply to conservative skepticism as follows:

> ED's response to conservative skepticism
> No we don't. We only need this kind of prior knowledge if there are epistemically motivated defeaters.

> Dogmatism's response to conservative skepticism
> No we don't. In the absence of defeaters, our experience as of p gives us prima facie justification for believing that p.

> Process reliabilism's response to conservative skepticism
> No we don't. We can know that p if our perceptual apparatus reliably produces true beliefs.

How can we characterize the dialectic between these three accounts and conservative skepticism? Process reliabilists, dogmatists and those who defend ED reject conservatism on the basis of liberal views concerning perceptual knowledge or knowledge in general. They do not provide any independent argument that conservatism is false for intrinsic reasons. Picking up Pritchard's terminology, they *override* conservatism about perception rather than *undercut* it. Thus, ED, process reliabilism and dogmatism might undercut skepticism understood as the general claim that we neither have perceptual knowledge nor knowledge that ~BIV. However, they only override *conservative* skepticism. Thus, the crucial point of dispute between ED and skepticism is the more general one between liberalism and conservatism, and the dialectic at that level can only be overriding, not undercutting. Again, ED and its rival accounts like dogmatism and process reliabilism are in the same boat.

Where does this leave us dialectically? I do not see any dialectical advantage for ED over alternative liberal accounts. Thus, ED is not the Holy Grail for the

skeptical problem. Either the Holy Grail is simply an account of basic knowledge, in which case any account allowing basic knowledge is an instance of it, or it is a conservative anti-skeptical strategy or the search has to go on.

Acknowledgements
The research was funded by the Austrian Science Fund (FWF): J 3174-G15.

References

Cohen, Stewart 2002: "Basic Knowledge and the Problem of Easy Knowledge". *Philosophy and Phenomenological Research* 65(2), 309–329.

Goldman, Alvin I. 1979: "What Is Justified Belief?". In: George Pappas (ed.), *Justification and Knowledge*. Dordrecht: Reidel, 1–23.

Huemer, Michael 2000: "Direct Realism and the Brain-in-a-Vat Argument". *Philosophy and Phenomenological Research* 61(2), 397–413.

Klein, Peter 2004: "Closure Matters: Academic Skepticism and Easy Knowledge". *Philosophical Issues* 14(1), 165–184.

Pryor, James 2000: "The Skeptic and the Dogmatist". *Noûs* 34(4), 517–549.

— 2004: "What's Wrong with Moore's Argument?". *Philosophical Issues* 14(1), 349–378.

Vogel, Jonathan 1990: "Are There Counterexamples to the Closure Principle?" In: Michael David Roth & Glenn Ross (eds.), *Doubting: Contemporary Perspectives on Skepticism*. Dordrecht: Kluwer, 13–29.

— 1990a: "Cartesian Skepticism and Inference to the Best Explanation." *Journal of Philosophy* 87, 658–666.

— 2000: "Reliabilism Leveled". *The Journal of Philosophy* 97(11), 602–623.

Grazer Philosophische Studien
90 (2014), 347–353.

Stefania CENTRONE (Hg.), *Versuche über Husserl*. Hamburg: Felix Meiner Verlag. 2013. 276 Seiten. ISBN 978-3-7873-2408-8, ISBN eBook: 978-3-7873-2409-5.

„Versuche über Husserl" werden im Titel des Buches bescheiden angekündigt. In der Hand hält man jedoch einen Band, der das Kunststück zuwege bringt, eine leicht verständliche und gut lesbare Einführung in Husserls zum Teil doch sehr komplizierte Gedankenwelt zu bieten und damit zugleich auch einen Einblick in die aktuelle Husserl-Forschung auf höchstem Niveau zu verbinden. Bei der Wahl der Themen hat die Herausgeberin eine ebenso glückliche Hand bewiesen wie bei der Auswahl der Autoren. Die Beiträge des Bandes beschäftigen sich mit den zentralen Themen der Husserl'schen Philosophie; sie reichen vom Problem des Psychologismus, der grundlegenden Bedeutung der Intentionalität für die Erkenntnis, der Rolle des Handelns für die Konstitution der Welt, den methodologischen Fragen der Reflexion, Deduktion, Eidetik und der Evidenz sowie der Frage nach dem Verstehen einer Person bis hin zur prinzipiellen Rechtfertigungsproblematik. Dabei werden auch die wichtigsten – direkten und indirekten – Gesprächspartner Husserls in die Diskussion einbezogen, nämlich Bernard Bolzano, Franz Brentano, Gottlob Frege, Martin Heidegger und Ludwig Wittgenstein.

Ein einführender Beitrag mit dem Titel „Leben, Werk und Wirkung" Husserls droht, eine staubtrockene Angelegenheit zu werden. Aus der Feder von *Wolfgang Künne* entsteht daraus eine überaus lebendige Geschichte mit vielen überaus interessanten Details, ohne Altbekanntes aufzuwärmen. Durch geschickte Auswahl und Zusammenstellung von Zitaten Husserls und seiner Schüler gestaltet Künne einen lebendigen Einblick in Husserls Leben, Werk und Wirkung. Dabei findet er zwischendurch auch u. a. noch Platz für eine wesentliche Richtigstellung: Husserls berühmtes Diktum „Philosophie als […] strenge Wissenschaft – *der Traum ist ausgeträumt*" war keineswegs (wie Ludwig Landgrebe annahm) eine Selbstkritik Husserls und ein Widerruf seines Programms der Philosophie als strenger Wissenschaft; es handelte sich dabei vielmehr (wie schon Gadamer festgestellt hatte) um den Ausdruck tiefer Enttäuschung darüber, dass mit dem durch den Nationalsozialismus proklamierten Ende der *autonomen Wissenschaft* auch eine Philosophie als *strenge Wissenschaft* unmöglich geworden war (21f.). Der scharfe Blick Künnes fördert aber auch Mängel zutage, die Anlass zum Schmunzeln geben; so z.B., wenn er aufdeckt, dass Husserls Bemerkung „Freges Kritik hat den Nagel auf den Kopf getroffen" mit „It hit the nail on *his* head" übersetzt wurde (Anm. 6 auf S. 24).

Künne gelingt es, auf dem knappen Raum von nur 15 Seiten die wichtigsten philosophischen Leistungen Husserls zu würdigen. Die kritische Auseinandersetzung mit diesen Leistungen ist nicht das Ziel dieses Beitrages, da dafür eine „Auseinandersetzung mit [Husserls] gesammelten Werken" erforderlich wäre; den „messianischen Anspruch, mit dem Husserl oft auftritt", kritisiert Künne aber dennoch als „befremdlich, wenn nicht verstiegen" (23).

Der Titel ‚Intentionalität' bezeichnet einen (wenn nicht *den*) Grundbegriff und das Hauptproblem von Brentanos deskriptiver Psychologie und von Husserls Phänomenologie. Diesem Thema widmet Wolfgang Künne seinen zweiten Beitrag in diesem Band (auf 97–143). Seine kritische Rekonstruktion der Husserl'schen Lehre erfolgt im Rahmen einer Gegenüberstellung mit Bolzanos Lehre von den Vorstellungen und Sätzen an sich.

In der Einleitung (97–102) zu diesem Beitrag schlägt Künne die begrifflichen Grundpfeiler für die systematische Behandlung der Intentionalitäts-Problematik ein: Intentionale Phänomene treten in zweierlei Form auf – als psychische *Akte* (z.B. Vorstellen oder Urteilen) oder als psychische *Zustände* (z.B. Lieben oder Glauben); daher ist die übliche Rede von intentionalen *Erlebnissen* nicht ganz passend. Außerdem können intentionale Phänomene *nominal* sein (z.B. Vorstellen oder Lieben) oder aber *propositional* (z.B. Urteilen oder Glauben). In jedem intentionalen

Phänomen steckt ein Gehalt; diesen Gehalt nennt Künne ‚Concept' (wenn es sich um ein *nominales* intentionales Phänomen handelt) oder ‚Proposition' (wenn es sich um ein *propositionales* intentionales Phänomen handelt).

Was ist nun aber unter der Intentionalität eines intentionalen Phänomens genau zu verstehen? Sie besagt, dass jedes intentionale Phänomen (und nach Brentano ist jedes psychische Phänomen intentional) auf einen Gegenstand „gerichtet" ist. Was aber heißt das? Dieser Frage widmet Künne den ersten Teil (§ 1) seines Beitrages (102–120). Für die nominalen Phänomene hat Bolzano wohl ein für allemal Klarheit geschaffen: Jeder Vorstellungsakt „*enthält*" ein Concept (bei Bolzano: eine Vorstellung an sich), das jedoch in vielen Fällen „ins Leere geht" und dem infolgedessen in diesen Fällen kein Gegenstand entspricht, entweder per Zufall (wie beim „goldenen Berg"), aus naturgesetzlichen Gründen (wie beim „perpetuum mobile"), aus begrifflichen Gründen (wie beim „hölzernen Eisen" oder dem „viereckigen Kreis") oder aus logischen Gründen (wie beim „nicht-viereckigen Viereck"). Mit diesem offenkundigen Faktum muss eine Theorie zurande kommen, die daran festhält, dass jede Vorstellung (oder gar jedes psychische Phänomen überhaupt) auf einen Gegenstand *gerichtet ist* oder *sich* auf einen Gegenstand *bezieht*. Solche und ähnliche Formulierungen erfordern interpretatorische Kreativität: Beim intentionalen Gerichtetsein handelt es sich (so besagt *eine* Interpretation) nicht um eine echte Relation, welche (wie viele Philosophen seit Aristoteles angenommen haben) die Existenz beider Relata voraussetzt, sondern bloß um etwas *Relativliches*, für welches die Existenz *eines* Relatums allein schon genügt. Oder aber eine *andere* Interpretation: Der Gegenstand, auf den ein psychisches Phänomen intentional gerichtet ist, muss kein existierender Gegenstand sein, sondern es kann sich dabei auch um einen *nicht-existierenden*, eben *bloß intentionalen Gegenstand* handeln.

Der ersten dieser beiden Interpretationen entspricht die Semantik für eine positive freie Logik mit einer partialen Interpretationsfunktion, der zweiten Interpretation entspricht eine Semantik für eine positive freie Logik mit einem *Outer Domain*. Bolzanos Auffassung hingegen entspricht die Semantik der negativen freien Logik, wie Künne richtig bemerkt (110): Wenn im Satz ‚die Vorstellung *v* stellt *N* vor' bzw. ‚*N* ist Gegenstand von *v*' anstelle von ‚*N*' ein leerer allgemeiner oder singulärer Name steht (wie ‚Landeinhorn' oder ‚Pegasus'), ist der Satz – im Rahmen dieser Semantik – schlicht und einfach falsch.

Neben diesen Deutungsvarianten gibt es immer noch die Möglichkeit, den an der sprachlichen Oberfläche als Nominalphrase auftretenden Intentionsausdruck in eine propositionale Phrase umzudeuten. Während die Deutungsmöglichkeiten für die (wirklich) *nominalen* Intentionsphrasen weitgehend systematisch ausgelotet zu sein scheinen, trifft dies auf die *propositionalen* Intentionsphrasen noch keineswegs zu. Das beginnt schon mit der notorischen Zweideutigkeit von dass-Sätzen (117): Mit einem dass-Satz kann nämlich sowohl sein Sinn (also eine Proposition) als auch der durch ihn beschriebene Sachverhalt gemeint sein. (Für Propositionen hat sich die Schreibweise mit eckigen Klammern, zwischen denen der auf das ‚dass' folgende Satz steht, eingebürgert; für Sachverhalte hingegen hat sich leider noch keine entsprechende Notation durchgesetzt – naheliegend wären dafür doppelte eckige Klammern.)

Jedes propositionale intentionale Phänomen „enthält" eine Proposition; die entscheidende Frage, die Künne aufwirft (116 ff.), lautet aber: Ist jedes propositionale intentionale Phänomen auch auf einen Gegenstand – nämlich einen Sachverhalt – gerichtet, und falls ja, in welchem Sinn? Künne plädiert bei den propositionalen Phänomenen für eine Lösung analog zu derjenigen bei den nominalen Phänomenen: Sofern man nichtexistierende Gegenstände ablehnt, muss es gegenstandslose nominale Phänomene geben; genauso muss es – aus realistischer Sicht – aber auch gegenstandslose propositionale Phänomene geben, jedenfalls dann, wenn man nicht-bestehende Sachverhalte ablehnt und nur bestehende Sachverhalte – also Tatsachen – akzeptiert. Nur einem wahren Urteil und einer wahren Überzeugung entspricht ein Sachverhalt, nämlich eine Tatsache, während den falschen Urteilen und Überzeugungen keine Gegenstände – wie nicht-bestehende Sachverhalte – entsprechen, da es diese (aus

der Sicht eines Realisten) gar nicht gibt. Diese Sichtweise haben u.a. Adolf Reinach und Anton Marty vertreten; Husserl hingegen war in diesem Punkt schwankend und oft unklar; Meinong wiederum hat jede Ambiguität, die sich in diesem Kontext auftut, in seine Objektive aufgenommen: Sie spielen zugleich die Rolle von Wahrheitsträgern und von Wahrmachern; es gibt wahre und auch falsche Objektive, und Objektive können (als Tatsachen) bestehen oder auch nicht bestehen.

Husserls Erklärung des – bei Bolzano und Frege ungeklärt gebliebenen – Verhältnisses eines intentionalen Phänomens zu seinem Gehalt – d. i. einer Bedeutungseinheit – bildet das Thema des zweiten Teils (§ 2) dieses Beitrages von Künne. In Künnes Terminologie ist dieser Gehalt eines intentionalen Phänomens (wie schon erwähnt) entweder ein Concept oder eine Proposition. Husserl fasste Concepte und Propositionen als Spezies von individuellen Momenten intentionaler Phänomene auf. Diese originelle Auffassung hat Husserl in den *Logischen Untersuchungen* entwickelt, schon bald darauf jedoch wieder verworfen. Künne verteidigt sie (und mit ihr den frühen Husserl) gegen Husserls spätere Wende. Dabei geht es Künne vor allem um die Aufklärung der Beziehung eines intentionalen Phänomens zu seinem Gehalt. Wenn man den ontologischen Status dieses Gehaltes – also eines Conceptes oder einer Proposition – in den Mittelpunkt der Betrachtung rückt, sieht es jedoch anders aus (jedenfalls wenn man dabei den Auffassungen von Bolzano und Frege gerecht werden will): Ein individuelles Moment eines intentionalen Phänomens kann nämlich nur existieren, wenn auch dessen Träger existiert; und eine Spezies kann nur existieren, wenn zumindest ein Exemplar von ihr existieren *kann*. Von entscheidender Bedeutung ist es nun aber, wie man dieses *Können* versteht: Wäre die Existenz der Spezies (also des Concepts bzw. der Proposition) nämlich von einer realen Möglichkeit eines Exemplars abhängig, ginge damit die von Bolzano und Frege hoch und heilig gehaltene Unabhängigkeit der Concepte und Propositionen von jeder Art von Bewusstsein verloren (wobei der katholische Priester Bolzano dabei speziell auch die Unabhängigkeit von den Ideen Gottes betont). Künne interpretiert dieses Können daher bloß als logische Möglichkeit bzw. Widerspruchsfreiheit (127); damit nimmt er allerdings in Kauf, dass unsere Ontologie ziemlich dicht bevölkert ist.

Stefania Centrone behandelt in ihrem Beitrag (65–96) die auch heute noch nicht ganz „erledigte" Problematik des logischen Psychologismus. Dieser trat historisch in zwei Varianten auf: als Begriffspsychologimus und als Gesetzespsychologismus. Stefania Centrone widmet jeder dieser beiden Formen des Psychologismus in ihrem Beitrag einen eigenen Paragraphen. Für die Problematik des Begriffspsychologismus dient ihr der Begriff der Anzahl als Illustrationsbeispiel. Dazu drängen sich Frege und Husserl als Protagonisten mit ihren Frühschriften (*Die Grundlagen der Arithmetik* von 1884 und *Philosophie der Arithmetik* von 1891) förmlich auf. In dieser Auseinandersetzung geht es aber schließlich und endlich nicht bloß um die Frage, ob eine psychologistische Definition des Anzahlbegriffs durch Abstraktion aus psychologischen Begriffen gerechtfertigt werden kann, sondern um die viel allgemeinere Frage, ob der Begriff der Anzahl überhaupt mit Hilfe einer sogenannten Definition durch Abstraktion bestimmt werden kann. In ein klassisches Beispiel eingekleidet, lautet die Frage: Ist auf einem festlich gedeckten Tisch die Anzahl der Messer *deshalb* identisch mit der Anzahl der Gabeln, *weil* auf dem Tisch *gleichviel* Messer wie Gabeln liegen, oder verhält es sich gerade umgekehrt: Es liegen auf dem Tisch *deshalb* gleichviel Messer wie Gabeln, *weil* deren *Anzahl identisch* ist? Mit gutem Grund stößt die Autorin vom Problem des Begriffs-Psychologismus zu dieser wesentlich grundlegenderen Frage und schließlich bis zur „Paradoxie der Analyse" vor.

Husserls Darstellung in den *Prolegomena zur reinen Logik* (= *Logische Untersuchungen*, Bd. I) bildet für Stefania Centrone den Ausgangspunkt ihrer Auseinandersetzung mit dem Psychologismus der logischen Gesetze. Extreme Psychologisten betrachten die logischen Gesetze als „Naturgesetze des Denkens" (Theodor Lipps), d.s. induktive Verallgemeinerungen von empirischen Beschreibungen realer Denkprozesse. Nach der Auffassung gemäßigter Psychologisten (wie Sigwart oder

Wundt) hingegen sind die logischen Gesetze keine empirischen Gesetze des Denkens, sondern sie sind *normativer* Art; sie beziehen sich zwar ebenfalls auf konkrete Denkprozesse, aber eben nicht rein beschreibend, sondern normierend. Sie sagen uns, wie wir denken und schließen *sollen*, wenn wir korrekt denken und schließen *wollen*. Warum aber *soll* man so-und-so und nicht anders denken und schließen? Weil diese Art des Denkens und Schließens (zum Unterschied von jener anderen Art des Denkens und Schließens) dessen Korrektheit und Gültigkeit garantiert. Eine solche Garantie können aber nur die *theoretischen Gesetze* der Logik liefern; diese sind jedoch nicht von empirischer, sondern von *apriorischer* Natur.

Dagfinn Føllesdal hat bereits in einer frühen Publikation (*Husserl und Frege*, Oslo 1958) grundlegende Gemeinsamkeiten in den scheinbar weit auseinander klaffenden Ansichten Freges und Husserls herausgearbeitet. Seither ist er zum wichtigsten Brückenbauer zwischen der Analytischen Philosophie (in ihrer ganzen Bandbreite) einerseits und den unterschiedlichen Varianten der „Weniger-bis-gar-nicht-analytischen Philosophie" geworden. Die Philosophie Husserls dient ihm dabei immer wieder als Verbindungsglied. Dass er gleich zwei Beiträge zum vorliegenden Band beigesteuert hat, ist ein besonderer Glücksfall. In dem einen der beiden Beiträge stellt er eine Verbindung von Husserl zu Heidegger und im anderen zu Wittgenstein her.

Die Rede von der *Konstitution* von Begriffen, von Gegenständen und letztlich von der Welt gehört nicht nur zum grundlegenden Vokabular der Phänomenologie, sondern wird genauso auch im logischen Empirismus (etwa in Carnaps *Logischem Aufbau der Welt*) und im Holismus von Quine verwendet. Während für Husserl die Konstitution der Welt primär eine Sache des Bewusstseins und seiner Intentionalität ist, verlagert Heidegger den Akzent vom theoretischen auf den praktischen Bezug zur Welt. Für Heidegger ist die Gesamtheit menschlichen Handelns an der Konstitution der Welt beteiligt. Husserl hat bekanntlich Brentanos Auffassung von Intentionalität dahingehend modifiziert, dass er zwischen den psychischen Akt und dessen intendierten Gegenstand das Noema zwischenschaltet: Das Noema bestimmt, ob es überhaupt einen Gegenstand geben kann, auf den der Akt gerichtet ist, oder nicht; und falls es einen solchen Gegenstand geben kann, wird er vom Noema eindeutig bestimmt. Das Noema spielt also die entscheidende Rolle in Husserls Konstitution der Gegenstände durch psychische Akte: sie sind *durch* ihr Noema auf einen Gegenstand gerichtet (sofern es einen solchen überhaupt gibt).

Die Übereinstimmungen zwischen Heidegger'schen Ausführungen in *Sein und Zeit* und Husserls Phänomenologie, auf die bereits Husserl in diversen Randnotizen seines Exemplars von *Sein und Zeit* hingewiesen hat, werden von Føllesdal detailliert herausgearbeitet. Er gibt sogar einen Übersetzungsschlüssel an, um zu zeigen, dass sich manche Stellen von *Sein und Zeit* tatsächlich wie eine Übersetzung von phänomenologischen Lehren Husserls lesen lassen. In den verschiedenen Weisen des In-der-Welt-Sein des Daseins wird nach Heidegger dieses Dasein selbst konstituiert; gleichzeitig wird aber umgekehrt auch die Welt durch das In-der-Welt-Sein des Daseins konstituiert (151). Dabei räumt Heidegger dem praktischen Bezug zur Welt den Vorrang gegenüber dem rein theoretischen Bezug ein (152 f.).

Wenn man nur die publizierten Arbeiten Husserls in Betracht zieht, geht Heidegger mit seiner Einbeziehung der Praxis (also von Handlungskomponenten) bei der Konstitution der Welt wesentlich über Husserl hinaus. Føllesdal, der Husserls Gesamtwerk wie kaum sonst jemand kennt, weist nun aber akribisch nach, dass auch für Husserl bei der Konstitution der Welt Handlungen eine wesentliche Rolle spielen und dass er damit viele Ideen Heideggers vorweggenommen hat (154–159). Nach Føllesdal bleibt jedoch offen, ob in dieser Frage Husserl mehr von Heidegger oder umgekehrt Heidegger mehr von Husserl beeinflusst wurde (160).

Diese Thematik wird im Beitrag von *Christian Beyer* über das Personenverstehen (255–276) wieder aufgenommen, ist doch das Personenverstehen „notwendig, wenn auch nicht hinreichend, für den erfolgreichen Vollzug eines sozialen Aktes" (263). Die kommunikative Umwelt und Lebenswelt und deren

Konstitution spielen dabei eine entscheidende Rolle (257–262).

Dem Thema der Rechtfertigung widmet Føllesdal seinen zweiten Beitrag (S. 167–192); dabei steht neben Husserls Auffassung diejenige von Wittgenstein im Mittelpunkt seiner Betrachtung. Føllesdal selbst hat in einer Reihe von Arbeiten die Methode des *reflective equilibrium* (was man im Deutschen für gewöhnlich mit „Überlegungsgleichgewicht" wiedergibt) verteidigt. Diese Methode wird meist mit den Arbeiten von John Rawls in Verbindung gebracht. Wichtige Vorarbeiten zu dieser Methode gehen auf Nelson Goodman zurück. Føllesdal führt Goodmans sogenannte Rechtfertigung der Deduktion (in *Fact, Fiction and Forecast*, Cambridge, MA, 1955) als Paradebeispiel für diese Methode an: „Wie rechtfertigt man eine Deduktion? Einfach dadurch, dass man zeigt, dass sie den allgemeinen Regeln des deduktiven Schließens entspricht […] Doch wie kann man entscheiden, ob Regeln gültig sind? […] Die Regeln des deduktiven Schließens werden gerechtfertigt durch ihre Übereinstimmung mit der anerkannten Praxis der Deduktion. Ihre Gültigkeit beruht auf der Übereinstimmung mit den speziellen deduktiven Schlüssen, die wir tatsächlich ziehen und anerkennen. Wenn eine Regel zu unannehmbaren Schlüssen führt, so lässt man sie als ungültig fallen. Die Rechtfertigung allgemeiner Regeln leitet sich also von Urteilen her, die einzelne deduktive Schlüsse verwerfen oder anerkennen. Das sieht eindeutig zirkulär aus […] Doch das ist ein guter Zirkel." (Nelson Goodman 1975: *Fact, Fiction, Forecast*. Deutsche Übersetzung: Frankfurt/M. Nachdruck 1989, 85ff.)

Ehrlich gesagt, kommt mir diese Art der Rechtfertigung höchst unplausibel vor. Die Sache ist doch viel einfacher: Wir rechtfertigen einen (als deduktiv gültig intendierten) Schluss, indem wir zeigen, dass er auf einer (korrekten) deduktiven Schlussregel beruht; und wir rechtfertigen eine (als deduktiv korrekt intendierte) Schlussregel, indem wir zeigen, dass sie uns nie von wahren Prämissen zu einer falschen Konklusion führen kann. In den seltenen Fällen, in denen ein Schluss lauter wahre Prämissen und eine falsche Konklusion hat, desavouiert sich dieser Schluss selbst als ungültig und damit als „unannehmbar"; und mit diesem konkreten Schluss werden auch alle anderen Schlüsse, welche dieselbe logische Form aufweisen, als „unannehmbar" verworfen. Auf der anderen Seite reichen jedoch auch noch so viele Einsetzungsinstanzen einer Schlussform mit mindestens einer falschen Prämisse und/oder einer wahren Konklusion nicht aus, um die Gültigkeit dieser Schlüsse und der entsprechenden Schlussform zu garantieren. Die Gültigkeit einer Schlussform bzw. der ihr entsprechenden Schlussregel, etwa des *Modus Ponens* ($A \rightarrow B, A \therefore B$), wird bekanntlich ohne jede Bezugnahme auf Einsetzungsinstanzen folgendermaßen bewiesen: $A \rightarrow B$ ist genau dann wahr, wenn A falsch ist oder B wahr ist; nun ist aber A wahr und infolgedessen nicht falsch; also muss B wahr sein. In diesem Beweis wird in der Beweis- bzw. Metasprache allerdings selbst wieder eine Schlussregel angewandt, nämlich die Regel des *Disjunktiven Syllogismus*, deren Gültigkeit für die Metasprache erst bewiesen werden müsste. Das führt zu einem unendlichen Rechtfertigungsregress bzw. zu einer unendlichen Rechtfertigungsspirale. Über kurz oder lang treten in dieser Spirale auch zirkelartige Phänomene auf, etwa bei der Rechtfertigung der *Simplifikationsregel* ($A \wedge B \therefore A$): $A \wedge B$ ist genau dann wahr, wenn A wahr ist **und** B wahr ist; wenn aber A wahr ist **und** B wahr ist, dann muss auch A wahr sein.

Auch wenn es sich hier nicht um einen echten Zirkel handelt (da es ja im einen Fall um eine Regel für den objektsprachlichen Junktor ‚$\wedge$' und das andere Mal um eine Regel für das metasprachliche ‚und' geht), macht dieses Vorgehen doch zumindest den Eindruck einer gewissen Zirkularität, da ‚$\wedge$' ein naher Verwandter von ‚und' ist. Operationalisten und Konstruktivisten haben daher gegenüber semantischen Rechtfertigungsstrategien immer wieder den Zirkularitätsvorwurf erhoben. Diese Vorwürfe wurden von semantischer Seite mit dem Hinweis darauf zurückgewiesen, dass es sich dabei ja nicht um echte Zirkel handle; außerdem sei die dabei zu Tage tretende Zirkelartigkeit harmlos. Eine solche Verharmlosungsstrategie ist jedoch gewiss nicht unproblematisch, lenkt sie doch von Problemen ab, denen man sich nicht entziehen sollte; wenn man in diesem Kontext

allerdings nicht bloß von einem harmlosen, sondern sogar – wie Goodman – von einem „*guten* Zirkel" spricht, so macht man damit ganz offenkundig aus der Not eine Tugend.

Die Pointe von Goodmans Rechtfertigung der Deduktion bleibt – zumindest für mich – im Dunkeln, so dass an dieser Stelle ein aufklärender Kommentar von Føllesdal sehr willkommen gewesen wäre. Umso wertvoller erweisen sich dafür Føllesdals eigene systematische Erläuterungen zur Rechtfertigungsmethode des Überlegungsgleichgewichts. Im Vordergrund stehen bei ihm vier Hauptmerkmale dieser Methode: 1. Kohärenz als Ziel; 2. ausnahmslose Revidierbarkeit aller Aussagen, inklusive der Beobachtungssätze (also umfassender Fallibilismus); 3. Anwendbarkeit in verschiedenen Bereichen (empirische Wissenschaften, Mathematik, Logik und Ethik); 4. intuitive Akzeptanz gewisser Aussagen vor aller Reflexion, wobei der Wahrnehmung ein gewisser Vorrang gebührt, obwohl auch sie (gemäß Punkt 2) nicht infallibel ist (176–181). Wenn die Methode des Überlegungsgleichgewichts auf alle vier (unter Punkt 3 angeführten) Anwendungsgebiete angewandt wird und es zwischen ihnen allen einen Rechtfertigungstransfer gibt, spricht Føllesdal von einem *globalen Holismus* (zu dem er sich selbst bekennt und den auch Morton White vertritt). Wenn jedoch in die Anwendung dieser Methode nicht alle vier Bereiche eingeschlossen werden, handelt es sich um einen *begrenzten Holismus* (wie demjenigen von Quine, bei dem die Ethik von der Anwendung dieser Methode ausgeschlossen bleibt, oder demjenigen von Duhem, der sich dabei auf die empirischen Wissenschaften beschränkte).

Bei den dabei erzielten überaus wertvollen Differenzierungen im Hinblick auf Rechtfertigungsfragen beruft sich zwar Føllesdal mehrfach auf Werke von Morton White, doch hat er selbst – als der wohl bedeutendste Interpret von Quines Holismus, den er in vielen Punkten kreativ weiterentwickelt hat – Entscheidendes zur Ausarbeitung dieser Methode beigetragen. (Überaus empfehlenswert zu dieser Thematik ist übrigens die gründliche Studie von Susanne Hahn: *Überlegungsgleichgewicht(e) – Prüfung einer Rechtfertigungsmetapher*, Karl Alber: Freiburg/München 2000.) Im vorliegenden Beitrag geht es Føllesdal vor allem darum, historische Belege für diese Rechtfertigungsmethode auch bei Husserl und Wittgenstein aufzuspüren. Bei Husserl sind dafür vor allem seine Überlegungen über die Lebenswelt relevant, die er z.B. in *Die Krisis der europäischen Wissenschaften* (1923/24, veröffentlicht 1976) und *Erfahrung und Urteil* (1939) entwickelt hat. Bei Wittgenstein findet Føllesdal – speziell in *Über Gewissheit* (1970) – gewisse Parallelen zur Methode des Überlegungsgleichgewichts in seinen Gedanken zum Begriff eines Systems und vor allem auch darin, dass er dem Weltbild eine wesentliche Rolle bei der Unterscheidung zwischen Wahrem und Falschem zuschreibt.

In seinem Beitrag über Husserl und Brentano beschäftigt sich *Markus Stepanians* – nach einer einleitenden Darstellung von Brentanos Leben und Werk und der Einordnung seiner Philosophie in die Geistesgeschichte des 19. Jahrhunderts – eingehend mit Husserls Kritik an Brentano. Im Zentrum von Husserls Kritik steht Brentanos Evidenzlehre: Während sich nach Brentano alle Urteile der inneren Wahrnehmung durch ihre Evidenz und damit Unfehlbarkeit auszeichnen, gibt es nach Husserl klare Beispiele für innere Wahrnehmungen, die nicht evident sind; in erster Linie kommen dafür Wahrnehmungen von psychischen Zuständen in Frage, die wir als „leiblich lokalisiert" empfinden, wie einen Schmerz *im Zahn* oder eine Angst, die einem *die Kehle zuschnürt* (54f.). Damit gibt Husserl ein wesentliches Fundament von Brentanos Philosophie preis.

Mit Husserls Auffassung von Evidenz beschäftigt sich auch *George Heffernan* in seinem Beitrag (219–254). Er kritisiert die reduktionistische Behandlung der Evidenz in *Die Idee der Phänomenologie* (1907). In einer akribischen Analyse weist der Autor nach, dass die von Husserl in dieser Schrift vorgenommene Identifizierung der Evidenz mit einer absoluten, adäquaten und apodiktischen Selbstgegebenheit mit seinem phänomenologischen Gesamtkonzept nicht oder zumindest weniger gut harmoniert als seine Beschreibungen der Evidenz in anderen Schriften (wie den *Logischen Untersuchungen* und der *Formalen und Transzendentalen Logik*); diese lassen nämlich auch eine relati-

ve, inadäquate und zweifelhafte Evidenz zu, die in *Die Idee der Phänomenologie* ausgeklammert wird (245). Daraus ergibt sich auch eine Relativierung der Verlässlichkeit dieser weit verbreiteten Schrift als grundlegender Einführungstext in Husserls Phänomenologie. Es handelt sich bei diesem Text jedenfalls nicht um „eine repräsentative Darstellung der Phänomenologie der Evidenz" (226; auch 246); dass er aber dennoch weitgehend als repräsentativ für Husserls Auffassung von Evidenz angesehen wird, erklärt dem Autor zufolge, warum Husserls Evidenzlehre und seine Erkenntnistheorie insgesamt in der analytischen Philosophie wenig Anklang gefunden haben und finden (247).

Eduard Marbach behandelt in seinem Überblick über verschiedene phänomenologische Methoden Husserls die *phänomenologische Reflexion* (198–202), die *phänomenologische Reduktion* (202–205) und die *eidetische Reduktion und Wesensanalyse* (205–213).

Stefania Centrone hat den Band überaus sorgfältig und leserfreundlich ediert und dabei insbesondere auch auf die einheitliche Gestaltung des Textapparates geachtet. Die Literaturverzeichnisse der einzelnen Beiträge wurden jeweils in Primär- und Sekundärliteratur unterteilt und untereinander abgestimmt, was bei einem Werk mit abgezählten 632 Anmerkungen keine Kleinigkeit darstellt. Damit gibt der Band ein kräftiges Lebenszeichen für die heute leider „vom Aussterben bedrohte Art" philosophischer Publikationen, in welchen jeder Interpretationsvorschlag durch Quellenbelege untermauert wird.

Edgar MORSCHER
Universität Salzburg

Grazer Philosophische Studien
90 (2014), 354–355.

Edgar MORSCHER, *Normenlogik. Grundlagen – Systeme – Anwendungen*. Paderborn: mentis Verlag. 2012. 309 S. ISBN: 978-3-89785-784-1.

Es gibt eine Reihe hervorragender Lehrbücher der modernen Modallogik auf Englisch (es sei nur verwiesen auf die Bücher von B. F. Chellas, von G. E. Hughes & M. J. Cresswell sowie von J. W. Garson), aber kaum ein in inhaltlicher und didaktischer Hinsicht ähnlich hervorragendes Lehrbuch auf Deutsch. Dasselbe galt bis vor kurzem auch für die Normenlogik; diese Lücke wird jetzt durch ein Buch von Edgar Morscher mit dem Titel „*Normenlogik. Grundlagen – Systeme – Anwendungen*" geschlossen. Es handelt sich dabei um ein sowohl inhaltlich als auch didaktisch sehr sorgfältig und gründlich ausgearbeitetes Lehrbuch der modernen Normenlogik.

Im Mittelpunkt der Darstellung steht das modale Standardsystem SNL der Normenlogik (es wird in der Literatur meist mit ‚D' tituliert) mit einigen Varianten dieses Systems. Die Ableitungen in diesen axiomatischen Systemen weisen keine Besonderheiten auf; der Autor verzichtet daher mit Recht auf eigene Ableitungsübungen im System SNL, da dies den Adressaten des Buches, die ja im Umgang mit axiomatischen Systemen der elementaren Logik bereits vertraut sein sollten, nichts Neues bringen würde. Statt dessen führt der Autor behutsam die beweistheoretischen und modelltheoretischen Grundbegriffe für SNL ein, die sich mühelos auch auf andere logische Systeme übertragen lassen und daher für die Leserinnen und Leser auch in ihrer weiteren Beschäftigung mit logischen Themen von Nutzen sind. Dasselbe gilt auch für die Beweise der Korrektheit und Vollständigkeit von SNL, die der Autor beispielhaft vorführt, so dass sie von Anfängern leicht nachvollzogen und in weiterer Folge als Muster für andere metalogische Beweise dienen können.

Über diese „Grundausstattung" eines Normenlogik-Lehrbuches hinaus bietet Edgar Morschers *Normenlogik* u. a. noch folgende „Zugaben":

1. In vielen Logik-Lehrbüchern wird den Leserinnen und Lesern die in den logischen Systemen verwendete formale Sprache und die dazugehörige formale Semantik einfach nur so vorgesetzt. Ein Vorzug von Morschers Lehrbuch der Normenlogik besteht darin, dass es in zwei eigenen Kapiteln durch informelle Erörterungen auf die formale syntaktische und semantische Behandlung der Normen im Rahmen der Normenlogik vorbereitet: In Kapitel 3 (17–50) wird durch eine Reglementierung von alltagssprachlichen Normsätzen Verständnis dafür erzeugt, dass die heute im Rahmen eines logischen Systems übliche formale Wiedergabe von Normsätzen mit Hilfe von normativen Modalphrasen bzw. Satzoperatoren eine Reihe von syntaktischen und semantischen Vorteilen mit sich bringt und daher nicht auf einer willkürlichen Entscheidung beruht.

In Kapitel 4 (51–84) werden verschiedene metaethische Standpunkte bezüglich der Interpretation von Normsätzen informell dargestellt und gegeneinander abgewogen; der Autor macht dabei keinen Hehl daraus, dass er zu einem non-kognitivistischen Standpunkt tendiert, den er hauptsächlich durch triftige Einwände gegen mögliche Alternativen zu stützen versucht. Dabei ist es ihm aber ein besonderes Anliegen zu zeigen, dass ein solcher non-kognitivistischer Standpunkt (wonach – grob gesagt – Normsätze weder wahr noch falsch sind) durchaus mit einer Mögliche-Welten-Semantik vereinbar ist, die später (in Kapitel 7, 107–115) informell entwickelt und danach (in Kapitel 8, 117–123) systematisch entfaltet wird.

2. Während die informelle Erörterung der syntaktischen Struktur und der Interpretation von Normsätzen in den Kapiteln 3 und 4 dem didaktischen Zweck dient, die formallogische Normensprache und die dazugehörige Mögliche-Welten-Semantik aus dem alltagssprachlichen Verständnis heraus zu entwickeln und dadurch an die Alltagssprache anzubinden, wird in Kapitel 10 (143–166) die umgekehrte Richtung betrachtet, nämlich folgende Frage behandelt: Wie kann man von den symbolsprachlichen Formeln wieder

zu informellen alltagssprachlichen sowie fachsprachlichen (z. B. rechtlichen bzw. rechtstheoretischen oder moralischen bzw. ethischen) Formulierungen „zurückfinden"? Dabei geht es darum, wie sich die Ergebnisse, die man exakt im formalsprachlichen System beweisen kann, auf die Praxis des normativen Argumentierens in alltags- und fachsprachlichen Kontexten zumindest approximativ anwenden und dadurch für die jeweiligen Anwendungsgebiete nutzbar machen lassen. Ohne eine solche Anwendung bliebe die Normenlogik, die ja auch Teil der Angewandten Logik ist, ein reines Luftschloss. Die Anwendung von Ergebnissen, die für formallogische Systeme erzielt werden können, auf Problemstellungen, die in der Alltagssprache oder einer nicht-formalen Fachsprache formuliert sind, ist zwar eine mühsame Angelegenheit, die man sich hart erarbeiten muss, sich aber dennoch nicht ersparen darf.

3. In einem eigenen Kapitel (Kapitel 12, 185–202) wird aufgezeigt, wie das Standardsystem zu einem multimodalen System ergänzt und weiterentwickelt werden kann; dabei spielen neben den normativen auch alethische und epistemische, insbesondere aber handlungslogische Operatoren eine wichtige Rolle, z. B. Belnaps stit-Operator ‚see-to-it-that p' oder ‚dafür sorgen, dass p'. Im Rahmen solcher multimodaler Systeme lässt sich eine Reihe von Schwierigkeiten und Paradoxien auflösen.

4. Eine wichtige Rolle bei der Anwendung von normenlogischen Systemen im Bereich von Rechtsphilosophie und Ethik spielt das so genannte Sein-Sollen-Problem. Ihm ist Kapitel 13 (203–232) gewidmet. In den meisten bisherigen Beweisen der Dichotomiethese (d. i. der These, dass kein nicht-trivialer reiner Normsatz aus einer konsistenten Menge rein deskriptiver Sätze logisch folgt) lassen sich logische Defekte feststellen. Morscher bietet präzise Formulierungen der Dichotomiethese in einer semantischen und einer syntaktischen Version und liefert für sie einen sorgfältigen Beweis im einfachen System SNL, ergänzt durch Hinweise, wie er auf kompliziertere normenlogische Systeme übertragen werden kann.

5. Der Band schließt mit einem Überblick über die moderne Normenlogik in den letzten 100 Jahren (Kapitel 15, 247–280); dabei sät ausgerechnet Georg Henrik von Wright, der Begründer der modernen Normenlogik, Zweifel am Status der Normenlogik.

Trotz dieser Vorzüge gibt es Monita. So ist im Kapitel 8 bei einer Reihe von expliziten Definitionen die reibungslose Eliminierbarkeit nicht gewährleistet, sobald diese Definitionen im Sinne einer Definitionskette aufeinander aufbauen. Man erkennt das, wenn man z. B. auf die Klausel (i) der Definition eines SNL-Modells von p. 119 die Definition eines SNL-Rahmens von p. 117 anwendet. Es zeigt sich dann im Ergebnis dieser Anwendung, dass die beiden Variablen ‚W' und ‚R' doppelt gebunden sind; und weiters folgt wegen $X = \langle W, R\rangle$ und $X = \langle W, R, V\rangle$, dass $\langle W, R\rangle = \langle W, R, V\rangle$; ein geordnetes Paar ist aber nicht identisch mit einem geordneten Tripel. Zur Behebung dieser Probleme könnte man z. B. auf p. 117 und p. 119 die Definition eines SNL-Rahmens sowie diejenige eines SNL-Modells vereinfachen, um so eine reibungslose Eliminierbarkeit zu gewährleisten, indem man definiert, dass ein geordnetes Paar $\langle W, R\rangle$ genau dann ein SNL-Rahmen ist, wenn (i) $W \neq \emptyset$, (ii) $R \subseteq W \times W$ und (iii) R seriell auf W ist, und weiters, dass ein geordnetes Tripel $\langle W, R, V\rangle$ genau dann ein SNL-Modell ist, wenn (i) $\langle W, R\rangle$ ein SNL-Rahmen ist und (ii) V eine SNL-Bewertungsfunktion für $\langle W, R\rangle$ ist. Jedenfalls fehlt – ohne derartige Vereinfachungen – auf p. 122 im Definiens der beiden Definitionen eines SNL-Modells* jeweils eine Bedingung, nämlich: $X = \langle W, R, V, w\rangle$. Und bei der informellen Vorbereitung der Konstruktion (A) auf p. 216 müsste noch als weitere Bedingung hinzugefügt werden, dass w_3 zu denselben Welten w' in der Relation R_3 steht, zu denen w_2 in der Relation R_2 steht.

Hans-Peter LEEB
Salzburg

Grazer Philosophische Studien
90 (2014), 356–361.

Lisa HERZOG, *Inventing the Market: Smith, Hegel, and Political Theory*. Oxford: Oxford University Press. 2013. X + 185 pp. ISBN: 978-0-19-967417-6.

Lisa HERZOG and Axel HONNETH (eds.), *Der Wert des Marktes: Ein ökonomisch-philosophischer Diskurs vom 18. Jahrhundert bis zur Gegenwart*. Berlin: Suhrkamp. 2014. 670 pp. ISBN: 978-3-518-29665-3.

After the recent financial crisis, there could hardly be a timelier topic than the understanding of markets. Lisa Herzog's 2011 doctoral dissertation at the University of Oxford, on which the book *Inventing the Market* is based, came just in time to meet the widely felt desire to get to grips with markets. Herzog has been incredibly productive in the past few years; there is not only her *Inventing the Market* and *Der Wert des Marktes* [The Value of the Market]; she has also written the entry 'Markets' for the *Stanford Encyclopedia of Philosophy* (2013), several free-standing papers on social and economic philosophy, and *Freiheit gehört nicht nur den Reichen* [Freedom Not Just For The Rich] (Munich: C.H. Beck, 2013), a popular-scientific monograph. Furthermore, she has edited the book *Hegel's thought in Europe* (Houndsmill and Basingstoke: Palgrave Macmillan, 2013) and co-edits, with Axel Honneth, *Joseph Schumpeter: Schriften zur* Ökonomie *und Soziologie* [Joseph Schumpeter: Writings on Economy and Sociology] (Berlin: Suhrkamp, forthcoming). She seems to have hit a nerve.

When talking about markets we mean, according to Herzog, "the complex system in which people buy and sell, offering money, goods, labour, time, and abilities. We all participate in it, day by day, in our roles as workers, customers, or investors." (1) Instead of leaving the field to economists, Herzog makes the case for a philosophical examination of markets in the sphere of political philosophy. What "needs to be addressed is", for example, "the meaning of markets for our identities, for our understanding of justice, and for the ways in which we are free or unfree." (4) As the subtitle indicates, she refers to Adam Smith and Georg Wilhelm Friedrich Hegel in order to reconstruct our understanding of the market. She believes that our current implicit understandings of markets, market societies and the role of the individual in these are by and large shaped by either Smith's or Hegel's theory. In her words, "it pays to revisit the writings of those who thought about market society at its beginning, and invented the views of the market that still influence our lives, both as intellectual constructions and as institutions and practices that have flowed from them." (5)

The book has seven chapters. Following the introduction, in which Herzog justifies her method against the Skinnerean critique (chapter I), the first two substantial chapters provide statements of Smith's (chapter II) and Hegel's (chapter III) understandings of markets and market societies. These chapters are very brief (about 20 pages each) and will likely leave some readers puzzled. Whom are they written for? It is certainly impossible to outline the relevant background theories and the particular views on the market by Smith or Hegel with so little space for non-experts. Readers familiar with Smith and Hegel, in contrast, will hardly get new insights. Herzog's main point in these chapters is to show "that Smith's and Hegel's views of the market and its place in society are much more similar than is often assumed." (9) The role of these two chapters becomes clearer in the second part of the book, which consists of three chapters that discuss problems related to markets and the market society, namely the self in the market (chapter IV), justice in the market (chapter V), and freedom, freedoms, and the market (chapter VI). In these chapters, Herzog frequently refers back to the preceding outlines of Smith's and Hegel's thought and further elaborates on their ideas regarding identity, justice, and freedom. After these two merely interpretive and three mainly systematic chapters, Herzog closes the book with a chapter on 'the market in history', where she argues that "philosophers and economists can benefit from a more historically situated approach to economic phenomena, which

takes into account the many and variegated forms that markets can take on." (16)

In chapter II Herzog presents Smith's theory as being much richer than the common cliché of "Smith-the-advocate-of-laissez-faire" (27) of pure market economy and self-interest has it, which is hardly surprising for anyone remotely familiar with Smith. He did not only write *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776) but also, inter alia, *The Theory of Moral Sentiments* (1759). It has been a common mistake—one that, as Herzog sets out, occurred primarily in the German-speaking world—to see these two works as irreconcilable. She further argues that it is not plausible to interpret Smith purely secular. Instead "it is likely that he shared the views of many 18th-century deists who believed in the ability of human reason to discover central tenets of a natural religion." (23) "The assumption that the world—including human nature—has been created by a benevolent deity forms the bridge from Smith's 'empirical' description of human nature and society to his normative moral theory." (24) And this is important for a proper understanding of the notion of the *invisible hand*. The basic idea, central not only to Smith but to the Scottish Enlightenment in general, is that good consequences can be attained without good intentions. Smith employs this thought not only in the economic sphere, but also in morality. According to him—and well supported by modern moral psychology—men naturally care most for the people closest to them. We feel most sympathy for our friends and family; but also more for the uncared-for child in our neighbourhood than for the starving children at some place on another continent that we only hear about in the news. The point is that, when everyone follows her natural tendency to care for the people close to her, then, ideally, everyone is being cared-for. Without actually intending this good consequence, we can achieve it by simply following our natural tendencies. We thus do not need to organise the care for every human being. "Smith explicitly claims that 'the care of the universal happiness of all rational and sensible beings, is the business of God', whereas to man is 'allotted a much humbler department, but one much more suitable to the weakness of his powers, and to the narrowness of his comprehension', namely to look after his own interest and the well-being of his family and country." (26; references to the original texts can be found in Herzog's book and are not repeated here for lack of space.) But, from time to time, the moral sentiments need to be developed through the faculty of reason and through the idea of the impartial spectator, the latter being a kind of corrective that demands such institutions that "lead to good consequences for *everyone* concerned, without sacrificing the interests of some to the interests of others." (27f., her italics)

For Smith, the market society has two prongs, first strong institutions that safeguard legal equality and that work against the tendency of wealth to translate into legal and political privileges; second, the free market as a sphere of production and exchange. Through the "desire of bettering our condition" (31), the division of labour, and the price mechanism markets "take over a task of coordination which could never be accomplished by an individual human being or a government." (32) Within the *Wealth of Nations* the *invisible hand* has two functions; one is to maximize the national product; the other is to benefit all members of society, including the poor. Wealth, Smith believes, trickles down. He "can put so much weight on self-interest in the market precisely because he thinks that a central task that other writers ascribe to benevolence, namely to take care of the poor and propertyless, is fulfilled by the market process itself." (34) Besides Smith's optimism that a benevolent deity has created a world in which good consequences can largely be attained without good intentions, there are some restrictions to free markets: He mentioned himself that free markets only yield the good when the economy is growing and that they cannot work in some areas of the public sphere, such as education, infrastructure, and stable financial systems. What he did not bother about too much is that this mechanism only works as long as externalities do not occur and when there is no negative trade-off between efficiency and welfare. As Herzog sets out, he was also extremely optimistic with regards to people's rationality. (35f.)

Hegel's thought is even harder to outline in 20 pages. Herzog does her best to introduce some of his main concepts and to avoid most of his rather dubious metaphysical assumptions. Drawing mainly on his *Philosophy of Right* (1821) she focusses on the ideas of *Geist* and *Sittlichkeit*. In his practical philosophy Hegel aims at "exploring the conditions for actualizing rational freedom in the modern state"; it's an examination if and how existing institutions and practices promote freedom. (45) *Geist* can be seen as the mediation of individual and one-sided struggles for recognition. Two individuals can ultimately recognize each other as free and equal only in the sphere of *Geist*, where they become part of a larger unit, as in the examples of patriotism or friendship. (46) *Sittlichkeit*, as the description of "customary roles of individuals in the institutions of family, civil society, and the state", is the 'living and present' *Geist* 'as a world'; *Sittlichkeit* thus comprises such institutions and practices that embed "the most advanced conception of human freedom present in a historical period." (48) The market society is part of the *Sittlichkeit* and of *civil society* in particular. "For Hegel civil society includes the 'system of needs', the 'administration of justice', and the 'police and corporations'; it can thus be described as the market economy together with the institutions that make it possible and that grow out of it." (53) The point is that the market society (the 'system of needs') is not freestanding; and it cannot be, because, for Hegel, in sharp contrast to Smith, the working of self-interest in the market yields chaos and disorganization. This might be surprising since Hegel also sees something similar to the *invisible hand* at work. What he calls the 'dialectical advance' means that "each man in earning, producing, and enjoying on his own account is *eo ipso* producing and earning for the enjoyment of everyone else." (1821, § 199) "Nevertheless, Hegel's market is not the peaceful, self-adjusting mechanism that Smith had described. […] The market is a battlefield of everyone against everyone else—and hence the 'relict of the state of nature'—and of each against the common interests of the community." (54) The greatest problem seems to be the unpredictability of markets, for producers as well as for workers. "With individuals' interests given free rein, 'accidental caprices and subjective desires' put people at risk, and make the satisfaction of their needs a matter of luck." (55) Furthermore, the poor will not "enjoy the broader freedoms and especially the intellectual benefits of civil society." (1821, § 243) As Herzog puts it, the "Smithian vision that economic growth would expand the cake for all is absent from Hegel's view of the modern economy." (56) This more pessimistic picture explains why Hegel complements the 'system of needs' within the civil society with the 'administration of justice' and with the 'police and corporations'. Whereas the former is roughly the legal system, 'police' is broadly understood to secure safety as well as to realize the right of 'every single person' to 'livelihood and welfare' (1821, § 230); the 'corporations', i.e. "the professional associations of those who work in the same branch of industry" (57), are very important for Hegel. For their members the corporations provide a sphere of recognition besides the family. The corporations largely regulate the economy and also support their members in a fashion very similar to social insurance. In sum, "Hegel does not *really* want to leave the economy free, in a way that builds on economic growth through the accumulation of capital. His focus is more on the distribution of work and of the necessities of life, and on questions like the 'honour' of individuals in the corporations; he does *not* build on labour as a mobile factor." (58, her italics)

Having seen that there are some differences between Smith and Hegel but that these are not as significant as some might have thought, one wonders why Herzog chose these two philosophers in the first place. Sure, both are interesting thinkers in their own right, as the vast amount of literature Herzog draws on shows. But would it not have been more fruitful to compare thinkers that in fact—and not only in common perception—draw very different pictures of the market? This worry gets some support from the second book under review here, *Der Wert des Marktes*, which Herzog co-edited with Axel Honneth, with whom she has worked at the *Institut für Sozialforschung* in Frankfurt. This book is a collection of texts (all in German) about the market and

the economy more general, ranging from Bernard de Mandeville's obligatory 'Fable of the Bees' (1705) to Eric Olin Wright's 'Transforming Capitalism through real Utopias' (2013). The book is divided into three parts, Rechtfertigung [justification], Kritik [critique], and Vermittlung [roughly, mediation or synthesis]. Each part starts with an introduction. What is striking is that Herzog and Honneth seem to have a clear preference for the third part of the 'Hegelian procedure' (9), for the mediation over the justification and the critique. Not only occupies the mediation part with its more than 300 pages half the book, its introduction—co-authored by Herzog and Honneth; the other two introductions are single-authored—alone being a 25-page treatment of the view on the market and political philosophy the two editors seem to endorse.

In this collection of texts, Smith has his place—along with Mandeville, David Ricardo, Friedrich August von Hayek, Gary S. Becker, and Rose and Milton Friedman—in the part that is devoted to the justification of the market. Consistent with Herzog's reading of Smith, the book provides excerpts from the *Theory of Moral Sentiments* as well as from the *Wealth of Nations*. The critical part features texts from Louis Blanc, Karl Marx, Rosa Luxemburg, John Ruskin, Karl Polanyi, Gerald A. Cohen, and Michael Albert. Hegel is to be found in the third part—along with John Stuart Mill, Émile Durkheim, Amartya K. Sen, Samuel Bowles, Albert O. Hirschman, Jens Beckert, Albena Azmanova, John E. Roemer, and Eric O. Wright.

Skimming through this wealth of different views—some more philosophical and argumentative in character, others more political and appellative—one wonders again and again why Herzog has chosen Smith and Hegel for comparison. Smith is arguably the first to have provided something like a proper economic theory that also explains the market. But Hegel is far from being an obvious choice. Neither has he explicitly engaged in economic thought. He comes very much from a political philosophy perspective that, admittedly, takes into account economic elements. But the latter have certainly not been the centre of his attention—also, in Herzog's entry on 'Markets' in the *Stanford Encyclopedia*, Hegel is only mentioned once and not discussed at all. The obvious question is this: Why Hegel and not, say, Marx? The latter would have provided a very critical account of markets and the market society. His theory is in sharp opposition to Smith's. Herzog's reason for drawing on Hegel instead of Marx seems to be the assumption that we cannot learn as much from Marx as from Hegel for our understanding of markets in contemporary liberal societies. She does not treat Marx precisely because he points out the internal contradictions of market-driven capitalism and ultimately rejects a market society. "In Hegel's political theory, on the other hand, the problems and contradictions are clearly seen, but they are analysed as capable of containment in a well-ordered society. […] Smith and Hegel, despite the differences in their views, stand within a liberal tradition, broadly conceived, for which economic liberties are compatible with other kinds of liberties within a stable social whole. Analysing their thought, and in particular their more critical remarks, thus allows us to develop *internal* criticisms of the liberal tradition, which seek to reform and improve it, while sharing its fundamental commitments." (9 f., her italics) This is a fair choice. But given the importance of critical voices on markets in the past two centuries and the very critical attitude towards them in contemporary culture and academia—at least outside mainstream economics—this choice would have called for a more elaborated explanation or justification.

To see whether the comparison between Smith and Hegel really is informative, let us have a look at two of the systematic chapters of Herzog's *Inventing the Market*. Chapter IV takes up the debate between communitarians and liberals whether markets create unencumbered, atomistic individuals and discusses the 'the self' in market societies. The relation between the individual and society seems to be a promising issue to find deep disagreement between Smith and Hegel. The former is often regarded using, or even having invented, the paradigm of the individual as being atomistic and free from commitments. Hegel, in contrast, is the paradigmatic contextualist for whom individuals are deeply embed-

ded in social structures. Not surprising for anyone familiar with Smith, his understanding of mankind was way more elaborated. "To be human for Smith means to share other people's feelings through sympathy." (63) Humans mirror their emotions in others and only through this process develop self-consciousness and self-command. This is also the reason why Smith put so much weight on education. Throughout their whole life people strive to live in private 'circles of sympathy'. An atomistic impression only appears in the market sphere. Here "sovereign individuals encounter each other as equals and exchange goods and services, each one recognizing that the others also have something to offer and respecting them as potential trading partners. […] The Smithian individuals treat their ability to work as human capital, that is, something they have at their disposal and can sell in the market—human capital is something they *have*, not something they *are*." (70, her italics) Having human capital means that they can change into other branches as they see fit; individuals are not embedded in their particular branch or company to the same degree as they are embedded in their private 'circles of sympathy'.

Hegel's view is not dramatically different. The greatest difference is that he puts much more emphasis on the embeddedness of the individual in corporations. Once one has chosen to work in a particular profession, this connection becomes constitutive of the self. The individual now is and is regarded as, say, a merchant. Being a merchant is what one is, not only what one has decided to offer on the labour market. Hegel believed that it is very hard, probably impossible, for people to change their occupation; "the thought that those who become unemployed, for example as a result of technical progress, can find work elsewhere seems to be foreign to him. This is why for Hegel the labour market needs to be regulated by the corporations." (74) From this discussion Herzog goes on to distinguish different spheres—for instance, private and institutional—some of which might call for embeddedness, whereas others might well work without and highlights that "Smith and Hegel did not yet seem very concerned about pressures from the market on the private sphere"—very much unlike Marx and Engels who charge "capitalism with destroying the workers' families and with reducing the bourgeois family to an instrument of procreation. In the 20th century, it was maybe Karl Polanyi's vision of society as a mere 'accessory of the economic system' that most clearly expressed the fear that all private relations might be completely dominated by forces of the market." (82) Again, it seems as if Marx, or Polanyi for that matter, would have provided a more interesting point of comparison to Smith's view.

Chapter V deals with questions of justice in the market, especially with the question whether talk of 'desert' makes sense with regard to markets. Can it be said that a certain market outcome is just because it is deserved? It can in Smith's system because the market rewards virtuous behaviour. "The basic argument is that in markets the free decisions by a large number of individuals result in patterns that resemble the judgments of an impartial spectator, and that this impartial spectator makes judgments based on an idea of desert: he holds that persons deserve certain rewards in virtue of having behaved in certain ways." (89f.) "Smith assumes that when individuals enter the market, their natural moral sentiments are not overridden by the desire to maximize their material gains." (111) Herzog argues, not very surprisingly, that this assumption is overly optimistic and does not really fit today's globalized markets, to say the least. Her treatment of the question whether or not this means that we should give up the notion of desert—or justice more broadly—as applied to markets, remains somewhat unsatisfactory. Herzog basically claims that we should not give up on justice; when we do not believe that the market itself yields justice, we have to design rules and institutions that frame the market in a way that supresses some of its bad effects, which is, by and large, Hegel's view. This, as well as the following discussion of how to theorize justice and the market, is relatively trivial. Just a random example, the behaviour of firms: "More transparency and more consciousness among customers are crucial for making markets more just, and better oriented towards transactions that are beneficial for everyone involved. The

Smithian ideal of markets as rewarding the provision of goods by practising the bourgeois virtues can here play a heuristic function: it can help us to ask whether we think a company can deserve the profits it makes, or whether it might have illegitimately exercised power over others, or violated basic rules or morality in some other way." (116)

To conclude, Herzog's monograph *Inventing the Market* is well written and generally informative. Smith and Hegel are in themselves interesting thinkers worth discussion. The choice to discuss the two in order to inform our contemporary understanding of markets has its pros and cons. Both do have interesting views on markets. But concerning some of the issues discussed in the book a comparison between Smith and a thinker more critical of markets would have been more rewarding. The baseline of the treatment of Smith and Hegel is all too often that they actually do have very similar views. Another problem of focussing on Smith and Hegel is that they often do have very different aims of explanation. Where Smith primarily develops market mechanisms and only subordinately links these to other human faculties and social institutions, Hegel developed a full system of philosophy in the idealist tradition and treats markets only in passing, often merely hinting at arguments for his claims. From time to time these different focusses make it somewhat difficult to compare their views.

The collection *Der Wert des Marktes* makes for a good reading for everyone new to the topic. Readers who are more familiar with the field will likely find the chosen texts too basic or the selection of texts relatively conservative. But a few of the contemporary texts might be new even to these readers. Herzog's and Honneth's introductions to the book's three sections are very knowledgeable overviews. Taken together the two books are a valuable source for readers who are curious what markets are, what they could be, and how they relate to other issues in political philosophy.

Norbert PAULO
University of Salzburg

Reference

Hegel, G. W. F. (1942/1821): *Philosophy of Right*. Translated by T.M. Knox. Oxford: Clarendon Press.

Information for Contributors

GPS publishes articles on philosophical problems in every area, especially articles related to the analytic tradition. Each year at least two volumes are published, some of them as special issues with invited papers. Reviews are accepted only by invitation.

Manuscripts in German or English should be submitted electronically as an e-mail attachment, either in MS Word or in rtf format, prepared for anonymous reviewing (i.e. without the author's name and affiliation), together with an English abstract of 60–100 words. Footnotes should be kept to a minimum, and references should be incorporated into the text in (author, date, page) form. An alphabetical list of references should follow the text.

A submitted paper will normally be sent to a referee.

Authors are responsible for correcting proofs. Corrections that deviate from the text of the accepted manuscript can be tolerated only in exceptional cases. Authors will receive a free electronic offprint of their article.

Manuscripts should be sent to the following address:
martina.fuerst@uni-graz.at

Email Addresses
Editors
johannes.brandl@sbg.ac.at
marian.david@uni-graz.at
maria.reicher-marek@rwth-aachen.de
stubenberg.1@nd.edu
Managing Editor
martina.fuerst@uni-graz.at

GPS online
Further information about GPS (back volumes, electronic publication, etc.) is available at the publisher's web site: http://www.brill.com/gps.