Morality and Institutions – On the bounds of ethics and economics and the unification of social sciences.

Gerhard Minnameier

1. Introduction: The quest for morality from an economic point of view

Morality seems to be multifaceted and elusive. Some think about it in terms of an other-regarding perspective (the "moral point of view")[1], others as rules that govern social life in smaller and larger social contexts.[2] Some think morality consists basically in reasoned moral judgements (Kohlberg, 1981; Colby & Kohlberg, 1987; Darwall, 2005; Gert, 2005; Scanlon, 2014), others believe that moral reasoning is mostly an ex post rationalization of what has been decided emotionally (in particular Haidt, 2001; Greene, 2013). In fact, many hold, today, that if judgements matter at all, then these are intuitive judgements made in "system 1" mode, rather than based on "system 2" reasoning (e.g. Kahneman, 2011; Greene, 2013; see Stanovich 2012 for an overview of similar approaches). Whenever people decide instantaneously, they usually take the moral course of action (provided they are moral persons); however, when they reason deeply, they usually choose wisely in terms of their own self-interest (Rand, Greene & Nowak, 2012; see also Cappelletti, Goth & Ploner, 2011; Cornelissen, Dewitte & Warlop, 2011).

While the message before and after the turn of the century was that real humans had social preferences that had to be built into economic models (Fehr & Schmidt, 2006; Dhami, 2016), views have changed after it has become clear that social preferences are mostly crowded out by anonymity and social distance (Hoffman, McCabe & Smith,1996; Dana, Weber & Kuang, 2007; Andreoni & Bernheim, 2009) as well as by changes in choice sets that seem to shift focal points (List, 2007; Bardsley, 2008). Dana, Weber, and Kuang got to the heart of the ensuing change of mind by asserting that "(r)ather than having a preference for a fair outcome, people may conform to situational pressures to give in certain contexts, but may also try to exploit situational justifications for behaving selfishly" (2007, 69).

However, one might wonder, then, why moral rules exist at all. And the question also arises, what motivates those who exert stable moral behavior, in particular those, who maintain cooperation in a one-shot prisoners' dilemma (PD), which account for roughly 10 percent (Ledyard, 1995, p. 172). Is their behavior rationalizable in any way, or is it downright and persistently irrational?

This question is, or should be, of interest not only for those working in the field of social preferences, but also from a wider economic point of view. Ever since James Buchanan asked "What should economists do?" (1964), has it been crystal clear that (modern) economics is basically a science concerned with human cooperation, broadly considered, i.e. cooperation in terms of economic exchange, collaboration based on loyalty towards people and organizations, reliability and trustworthiness in business relationships and so forth.

---

[1] For instance, Frankena defines morality as "a normative system in which evaluative judgments … are made … from the point of view of a consideration of the effects of actions, motives, traits, etc. on the lives of persons or sentient beings as such, including the lives of others besides the person acting, being judged, or judging" (1980, p. 26). For Turiel morality consists in prescriptive judgments "about welfare, justice, and rights … that involve concern with dignity, worth, freedom, and treatment of persons" (2006, p. 10).

[2] Haidt speaks of morality as an "interlocking sets of values, virtues, norms, practices, identities, institutions, technologies, and evolved psychological mechanisms that work together to suppress or regulate self-interest and make cooperative societies possible" (Haidt, 2012, p. 270).

2. Morality in terms of social preferences

If we want to understand morality in terms of social preferences, we take on a basically decision-theoretic point of view, in which the other agents (with their perceived preferences and beliefs) are understood as the restrictions, under which the agent forms an intention. These restrictions can be positive (i.e. affordances) or negative (i.e. constraints). Intentions can be formed deliberately or intuitively. A rational moral choice would then be utility-maximizing in terms of the agent's underlying preferences and the restrictions under which she chooses.

On this view, we have to distinguish between two completely different notions of "preference". The first is the concept of "revealed preference" which pertains to consequences of choices (Samuelson, 1948). The second is the concept of underlying fundamental preferences that motivate concrete (revealed) preferences. For instance, a preference for small car, if a larger one were clearly affordable, may be motivated by thriftiness or by a concern for the environment (Dietrich & List, 2013a and b). In the moral domain, principles like the "Golden Rule" might be motivating reasons for the proposer in the dictator game to prefer a fair outcome, or for the proposer in a trust game to transfer a fair share back to the investor.

It is important to note that these fundamental preferences are out of reach for the strictly behavioristic concept of revealed preferences which typically relate to observable outcomes of choices. In this sense, rational choice theory based on "revealed preferences" has been criticized for depriving utility of all content (Hollis & Sugden, 1993; Bruni & Sugden, 2007). Conversely, the idea of fundamental preferences (re-)endows utility with content. It brings the central drivers of choices to the fore, among them also what we generally conceive as social preferences.

While this approach has much in common with the psychological approach originally set forth by Gary Becker (1976), it also differs importantly from it. Becker's notion of "basic preferences" refers to motivators common to all humans at all times, much like the basic needs assumed by Deci and Ryan in the self-determination theory of motivation (2000; 2017). In contrast to this, fundamental preferences may vary from inter-individually. Thus, specific moral abilities and orientations are captured by this notion of fundamental preferences, but not by basic preferences in Becker's sense.

Hence, we can understand and analyze moral agency in a rational-choice-theoretical frame of reference, where a moral person has fundamental preferences in terms of her deep moral convictions, but faces certain restrictions in a specific situation. Consequently, a rational choice (in the prescriptive sense) is one that best suits the moral goals of that person in that particular situation. And in terms of an analysis of human behavior, we can explain why one and the same person chooses differently in different situations, given that the restrictions are characteristically different.

A simple example may be in order. Let us imagine a person who tries to follow the "Golden Rule" ("Do unto others what you would have others do unto you"). Her default strategy in a PD might be to cooperate, since she would not wish her partner to defect. However, given the risk of defection and her inability to prevent it, this strategy is self-defeating, which explains why not only self-interested, but also morally motivated people learn to defect. If the latter defect, they do not do so for their own self-interest, but because they face restrictions they cannot ignore. In that very sense, Ellingsen et al. (2012) have revealed that social framing effects in the PD do not depend on frame-specific preferences, but on frame-specific believes. When in a so-called "Stockmarket Game" players consider the risk of defection higher than when in a "Community Game".

Of course, there are also individuals who keep on cooperating despite negative experiences (Ledyard, 1995), and the question arises whether their choices ought to be considered as irrational or, on the contrary, as the truly moral courses of action. I would like to postpone this question here and raise it again in section 7. However, given that a moral person believes that she ought to defect in a one-shot PD it would be irrational, if she cooperated none the less. And it would be equally irrational if she thought it worth a try to cooperate, but then defected. Thus, the reason-based theory of rational choice also allows

for irrationality (Dietrich & List, 2013a**).** This is another important difference compared with Becker's analytical model in which any behavior is "rationalized".

Reconstructing morality in terms of the reason-based theory of rational choice, opens two important perspectives for the theory of social preferences. *First*, preferences are not understood as (classes of) behavioral outcomes. In the tradition of revealed preferences, economists are sometimes too quick to assert preferences in terms of the consequences of acts, like e.g. choosing an equal result (which indicates inequity aversion) or an efficient outcome (revealing a efficiency orientation) or allotting the highest amount to the worst-off person (which expresses a maximin preference). In may well be that we have to re-analyze such outcome orientations in terms of downright moral principles that motivate these choices. Here, moral psychology (and moral philosophy) could inform behavioral economics and vice versa.

Second, a clear understanding of fundamental moral preferences may entail an equally clear understanding of the relevant conditions (restrictions), under which they (do not) work. The one-shot PD is a situation that suggests a specific kind of morality: that everyone follows their self-interest. To be sure, this is a moral judgement, indeed, because it pertains not only to the self, but also to the other player. As we have seen, however, the Golden Rule does not work in these circumstances.

In moral psychology it has long been clear that people follow different moral guidelines in different situations (Krebs & Denton, 2005; Beck, 2008; Rai & Fiske, 2011), because different situations seem to require (or afford) different kinds of morality. Thus, when economists ask the question "Which model of other-regarding preferences does best in the light of the data and which should be used in applications to economically important phenomena?" (Fehr & Schmidt, 2006, 668), it may not be well put. For the answer might be that different such preferences, or rather moralities, might me appropriate in different situations.

3. Morality in terms of institutions

However, understanding morality in terms fundamental preferences may still not be the entire truth. To see why, we have to recall the two different notions of morality already discussed above. So far we have considered moral principles in terms of other-regarding preferences, and thus have put them in a decision-theoretical frame of reference, in which the moral agent determines what she ought to do, given her (moral) preferences. Ken Binmore has criticized this approach for "blurring the distinction between a social norm and a social preference" (2010a, 141), and believes this "will turn out to be a bad mistake" (ibid.), basically because it reframes a game-theoretic problem as decision theoretic one.

Conversely, if we understand moral principles in terms of rules that govern social interaction, we take on a game-theoretic perspective. In this latter sense, moral principles are not to be understood (merely) as moral preferences, but as social institutions, i.e. as rules of the games that agents play.

This is the view, Binmore (2010b) advocates. However, he also raises the important question, whether social norms (or moral principles, as I prefer to call theme here) can really take on the role of rules of the game in the game-theoretic sense, because for this they have to be unbreakable. If moral principles are simply rules in terms of socially shared norms about what one ought to do in certain situations, they would imply that one could also violate them. However, if they can be violated (especially by self-interested people), they cannot possibly be rules of the game in the game-theoretic sense.

If one accepts this, which I do, moral rules have to be shown to be unbreakable, indeed. Moreover, if they are to function as institutions, they have to come with sanctions that enforce them and make cooperation the dominant strategy. This is perhaps an unfamiliar way of looking at morality, since many have claimed that moral rules motivate the invention and implementation of institutions (Brennan & Buchanan, 1985; Homann, 2016), but only very few have come to understand moral principles themselves as institutions or as social norms in the economic sense.

Cristina Bicchieri (2006; 2017) is among them, and she discusses norms as solutions to mixed-motive games in Schelling's (1960) sense. If she is right, morality can be assigned a proper and a prominent place in economics, because moral principles as a whole would then be the toolbox we use to make cooperation possible. And if Buchanan is right in claiming that economic problems are typically problems of human cooperation, the relevance of morality for economics can hardly be overestimated (see also Binmore, 2011, esp. p. 171**;** Luetge, Armbrüster & Müller, 2016).

Let us return to the question of how moral rules can be unbreakable, and let us consider the simple example of the Hawk-Dove Game (cf. Gintis, 2014, pp. 42-43) . In Hawk-Dove**,** players compete over a resource of value v and have the following two strategies: The first is to play "hawk" (H), which means to fight until one either wins the territory or is injured and has to retreat (with cost of injury c). The second is to play "dove" (D), which is to display hostility, but retreat before any kind of fighting would actually start. Let $v = 20, c = 40$, and let us further assume that both combatants have equal strength, so that there is a 50 percent chance of winning for each of them. Figure 1 shows the resulting strategy profiles and the payoffs.

	H	D
D	*0, v*	*v/2, v/2*
H	*(v-c)/2, (v-c)/2*	*v, 0*

A

	H	D
D	*0, 20*	*10, 10*
H	*-10, -10*	*20, 0*

B

Figure 1: The hawk-and-dove game with (a) the payoffs in general form and (b) the payoffs if $v = 20, c = 40$

Unlike the PD, HD does not have a stable pure strategy Nash equilibrium. There are two pure strategy Nash equilibria (H, D and D, H) which, however, are asymmetric. On top of these, there is a symmetric equilibrium in mixed strategies, in which each player chooses H and D with probability $p = .5$. In this case, each player earns an expected payoff of 5. However, similar to the PD, this symmetric mixed strategy Nash equilibrium is Pareto-inefficient, because both players end up with 5, when 10 would have been possible in (D, D).

Fortunately, there is a way out – not within the game, but by augmenting the game and thereby changing it. This is done by introducing a new rule, the so-called "property rule" (P). This new rule turns HD into a new game ("Hawk-Dove-Property", or HDP) (Gintis, 2014, pp. 145-146) The rule introduces a new strategy P, namely: "When first at the resource, play H, otherwise play D." This is a realistic and common rule which is used, for example, when people compete over seats on trains, parking spaces, spots for sunbathing on the beach, etc.[3] What's more, this is not just any rule, but clearly a moral rule. It establishes an elementary property right or a temporary right to dispose of something (e.g. a certain toy in a kindergarten).

Let us now look at this new game HDP, and let us further assume equal chances to be the first (or second) at the resource. Under these conditions, we obtain the following payoff matrix (see 2). For reasons of simplicity, only the row-player's payoffs are shown.

[3] Furthermore, it is also common for people when they come in second or so, that they still try to impose themselves in case the other party might be too gentle and compliant. Therefore, it seems quite realistic that such an agent plays D rather than backing out right away.

	H	D	P
P	*(v-c)/4*	*3v/4*	*v/2*
D	*0*	*v/2*	*v/4*
H	*(v-c)/2*	*V*	*3v/4 – c/4*

A

	H	D	P
P	*-5*	*15*	*10*
D	*0*	*10*	*5*
H	*-10*	*20*	*5*

B

Figure 2: The Hawk-Dove-Property game with the payoffs for the row-player: (a) in general form and (b) if $v = 20, c = 40$

If the column-player chooses P, the row-player's best response is to choose P, too. Hence the payoffs (10, 10) are now a Nash-equilibrium, which it was not in HD. And whereas HD has the structure of a "mixed-motive game" according to Schelling (1960), the introduction of P has transformed it into a coordination game. This seems to be the basic characteristic of all moral principles, that they transform mixed-motive games, in which Pareto-superior strategy combinations are not a Nash-equilibrium, into coordination games in which they are.

If this is correct, then moral principles function as rules of the game, indeed, because then they are unbreakable (at least on pain of irrationality). Of course, both players have to understand the rule and prefer to play the game like this. However, this is a basic game-theoretical requirement, since if things were different, the game itself would be different.

For instance, the game would already change, if the column-player decided to play strategy D. In this case, we see that this tantamount to an invitation for the row-player to choose H, since now H maximizes her payoff. The important lesson to be learnt from this is that if moral principles (or social norms) are institutions, then the respective sanctions have to be used. Here, the sanction consists in claiming one's property and being prepared to fight in case of violations. If (all too) good-hearted people refuse to play those sanctions, they are not playing a moral game anymore, but are corrupting morality and destroying the game (because rule P then ceases to be a rule of the game).

4. Correlated equilibria and institutional economics

Correlated equilibria, previously a concept of cooperative game theory, were introduced by Robert Aumann (1974) in the context of non-cooperative game-theory. From the point of view of cooperative game theory, they determine only what is beneficial for a certain coalition of agents, not how they might achieve it. However, as Gintis stresses, "the correlated equilibrium is a much more natural equilibrium criterion than the Nash equilibrium, because of a famous theorem of Aumann (1987), who showed that Bayesian rational agents in an epistemic game G with a common subjective prior play a correlated equilibrium of G" (Gintis, 2014, p. 142). This raises the question of how they manage to coordinate in this way. Of course, the straightforward answer is that they must do something to the effect, that the correlated equilibrium becomes also a Nash-equilibrium. As we have seen in the above example, this involves changing the original game, here by transforming it from "Hawk-Dove" to "Hawk-Dove-Property".

Such a move is not part of classical game theory. And it cannot even be part of game theory in a narrow sense, because it pertains to transforming or inventing games. Changing the (rules of the) game is never a strategy within the game. And even if the enhanced game is not creatively invented, but known by the players prior to implementing it, this implementation is something they have to do before they can actually play it. That is, they have to agree to play by the new rules. In this sense, the setting of institutions is not equivalent to playing a meta-game, but a process that transcends game theory as such.

Therefore, the task of setting new rules is technically assigned to a "choreographer" who implements them (Gintis, 2014, p. 44). However, we can also imagine the players themselves establishing new rules and subsequently playing according to them (Fudenberg & Tirole, 1991, 53). They might be motivated by the insight that the game puts them in some fatal competition which is disadvantageous for both sides. In other words, they face a cooperation problem, and once they realize that they are caught in it, they try to change the game.

In Hawk/Dove/Property they establish the institution of "property", which allows them to achieve the correlated equilibrium. However, within this new game, new cooperation problems may arise. Once property is established, it might happen that player A lacks what player B has and vice versa. Sharing and turn-taking are the simple institutions with which this problem can be solved. We end up with a more complex game that involves both the property rule and the sharing norm.

And, finally, another novel problem arises when the players are not the same. For instance, imagine one is rather rich and can afford many things, while the other is poor. Would we expect that the poor player to share in the same way as the rich one? No, we would rather expect the rich player to share and expect nothing in return. And if the rich were deprived in some sense or for some reason, she would also merit our care.

Another relevant difference between the players is a difference in effort or expertise. If two agents work together on a joint project and one puts in much higher effort than the other or has a higher expertise (say one is the expert, the other a novice or helpmate), then the returns from collaboration should not be shared equally, but equitably. Thus, if we have relevant differences between the agents, the principles of care and equity ought to be applied.

What we see, is a cascade of games, one built on top of the other. And with every new problem, a new kind of morality emerges that functions as an institution to make correlated equilibria achievable. In other words, we see a cascade of mixed-motive games that are turned, one after the other, in coordination games. How is this possible?

5. Moral principles as institutions

For moral principles to function as institutions, they have to go with sanctions, so that the payoffs be changed (to the effect that the game ceases to be a mixed-motive game and is turned into a coordination game, like in the case of Hawk/Dove/Property). What sanctions might these be? – These sanctions are, of course, subtle, and as Bicchieri (2006, pp. 42-46) has pointed out, we are so used to playing these games and have internalized their benefits to such an extent, that we play them as if they were coordination games. However, we can make very clear, none the less, what sanctions apply and how they work.

The examples I have mentioned above pertain to simple principles of morality that typically apply to close relationships, where one feels affiliated with the other players. At least, let us consider them first in these circumstances. In such situations the players have an interest in each other and in each other's well-being, and therefore the moral currency that applies are "liking" and "dislike". To be sure, we not only use negative sanctions (punishments) but also positive ones (rewards). Whenever somebody shares with us, we answer this by thanking the person or sending other signals of approval. If the person does not share, we do the contrary.

Moreover, if someone repeatedly fails to enact the respective morality by failing to share in situation where one should share, the ultimate move is to stop the game and revert to the lower-stage form. In the present case this means to revert to a game, in which nobody shares, but where everybody claims their property rights. In other words: If you fail to share with me, I will stop sharing with you, and we stay with what we have. This can be understood as an extended form of punishment, which can also be used as a threat as long as one still is within a game of mutual sharing. If other players even violate my property, I could first send signs get angry at them and express my indignation, but the ultimate move

could be to start conquering back items from others. The latter would be tantamount to reverting to a Hobbesian state of nature, in which, according to the "law of Nature (…) every man has right to everything" (1651/2001, p. 65 [Chap. 15, §2]), but which results in a "war of every one against every one" (1651/2001, p. 59 [Chap. 14, §4]).

However, as long as the respective moralities function, cooperation is upheld, because respecting each others property or legitimate rights of use, being prepared to share with each other, taking care of the other seeking equitable exchanges, all these attitudes and behaviors create and promote friendship and a mutual concern and sympathy for each other.

The three forms discussed so far have one thing in common: the fact that they build on (mutual) sympathy, i.e. a concern for the other's wellbeing and a concern to gain and uphold the others liking or even friendship. Thus, they apply to close relationships, but not to interactions in the wider social world.

In the tradition of Lawrence Kohlberg's theory of moral stages, I refer to this sympathy-based morality as Stage 1, with three substages A, B, and C, as shown in Table 1.[4] The sequence of substages follows a Piagetian developmental logic, according to which the perspectives of different individuals are first differentiated (A), then reciprocally related (B), and finally integrated (C). The basic property right as in Hawk and Dove, for instance, allows a person to coordinate her perspective with that of another person, because it equally applies to self and other. Both perspective are taken independently from each other, and the principle applies equally to self and other. The sharing norm, in turn, establishes a reciprocal relation that regulates transfers between the two sides. These transfers, however, are independent from relevant inter-individual differences, which are only integrated at the C-substage in terms of care and equity. There is also relevant evidence on this from pre-school children's sharing behavior (see e.g. Paulus & Moore, 2014).

Table 1: Neo-Kohlbergian Stage 1 with substages A, B, and C

Stage	Principle	Reward	Punishment	Ext'd punishment (one stage down)
1A	Property	Liking	Dislike	Revenge (→ state of nature)
1B	Sharing/turn taking	Liking	Dislike	Stop sharing (→ property)
1C	Care/equity	Liking	Dislike	Strict reciprocity (→ sharing/turn taking)

These three moral principles discussed so far all rely on sympathy and a harmonious mediation between self and other (see also Paulus & Moore, 2017). However, there are also true conflicts of interest, in which helping others may be fatal. Consider, for example, the situation of a couple of graduates applying for jobs, over which they (have to) compete. Here, they may even feel a lot of sympathy for each other; yet, they clearly are competitors. In such a conflict of interest the agents have to follow the rule that everybody has their own (legitimate) interest to pursue, or in a proverb: *"Near is my shirt, but nearer is my skin"*. This is not equivalent with selfishness, because one also respects that others do the same (or

[4] First, note that Kohlberg first introduced „sub-stages" A and B (see e.g. 1984), but later treated them as mere „types", because he noticed anomalies in the developmental sequence (Colby & Kohlberg, 1987). From the point of view of the neo-Kohlbergian taxonomy, however, there are not anomalies, so that one can rightly speak of sub-stages A, B, and C in the neo-Kohlbergian framework. Second, although Kohlberg has not referred to sympathy as basis for reasoning at his Stage 1, it may be thought to be immanent in his view that at Stage 1, "moral judgments are self-evident, requiring little or no justification" (Colby & Kohlberg, 1987, p. 25). His examples are in part examples of neo-Kohlbergian Stage 1A, e.g. that "telling on your brother is wrong because that is tattling and breaking into the druggists' is wrong because 'your are not supposed to steal'"(ibid.). The first implies sympathy with the brother and the negative consequences he might face, if one snitched on him, the second is an example of a property right, as explained in the text. When Colby and Kohlberg write that "distributive and retributive justice are characterized by strict equality rather than equity" (ibid.), this exemplifies Stage 1B in the neo-Kohlbergian framework. Conversely, neo-Kohlbergian Stage 1C would already be scored at Kohlberg's Stage 2, on which he explains: "There is reference to individual needs or intentions as the basis for equity rather than strict equality or literal reciprocity in distributive justice" (ibid., p. 26).

in fact, have to do the same). Furthermore, in some cases one can mitigate the conflict by throwing dice or so. In this case, the winner is determined in a fair and easy way. However, it still is a conflict of interest and it still involves respect for the winner and for the procedure. Thus, *respect* is the currency, in which the sanctions are valued in conflicts of interest.

This morality is labelled Stage 2A.[5] It differs from Stage 1A in that not necessarily everybody has his or hers (i.e. one's own property or things to dispose of). At Stage 1 one's own interest and the interests of others are necessarily compatible – at least as far as one's sympathy for the other goes. At Stage 2, however, there is a real conflict of interest, which also applies to the following substages 2B and 2C.

Stage 2B relates to conflicts of interest that cannot be addressed by having everybody pursue their own interest, as at Stage 2A, since it refers to situations where this orientation entails a social dilemma. The classical PD is a case in point. Of course, in a strict PD the only option is to follow the principle of Stage 2A, because it forces players to follow their personal self-interest. How, under more relaxed conditions, where communication is possible, the morality of contract allows agents to solve this problem and attain the Pareto-superior state. Again, the tools to enforce a contract that prevents both prisoners from confessing consist in the positive and negative sanctions with which the contract is enforced (e.g. a credible threat that your buddies would avenge you, if the other confessed, and the prospect of trustful collaboration in the future, if the promise is kept).

Finally, Stage 2C pertains to the Golden Rule. Here, one does not strike a deal with the other, as with Stage 2B, but one rather makes a deal with oneself in terms what one would endorse if one were the other. This applies to cases in which the other has nothing to offer and therefore cannot strike a deal in any way. A trust game (Berg & Dickhaut, 1994; Johnson & Mislin, 2011) incorporates this idea. To play it wisely requires that both players understand this morality and that the investor has reason to believe that the proposer would play according the Golden Rule. Only then can this moral game be played and can the players obtain a fair and efficient payoff.

Table 2: Moral principles and sanctioning mechanisms for Stages 1 and 2

Stage	Principle	Reward	Punishment	Ext'd punishment (one stage down)
1A	Property	Liking	Dislike	Revenge (→ 0)
1B	Sharing/turn taking	Liking	Dislike	Stop sharing (→ 1A)
1C	Care/equity	Liking	Dislike	Strict reciprocity (→ 1B)
2A	Legitimate Interest	Respect	Disrespect	Suspension/separation (→ 1C)
2B	Promise/contract	Respect	Disrespect	Defection in PD/self interest (→ 2A)
2C	Golden rule	Respect	Disrespect	Tit-for-tat in PD (→ 2B)

So far, two elementary moral stages and their three substages have been described. In my overall taxonomy, there are nine stages altogether which, of course, cannot be explained and illustrated in the paper (see Minnameier, 2000, for a comprehensive treatment, and Minnameier, forthcoming, for an extensive explanation of Stages 1 through 3). However, the basic idea to be conveyed here is that moral principles function as solution concepts for mixed-motive games that are transformed as a consequence into coordination games. Furthermore, these games form a hierarchical order, so that each morality based coordination game that has evolved out of a mixed-motive game ultimately leads into a new mixed –motive game at a higher order. Hence the hierarchy of moral stages and the hierarchy of games.

[5] Note that this would also conform to Kohlberg's understanding of Stage 2: "There is an awareness that each person has interests to pursue and that these may conflict" (Colby & Kohlberg, 1987, p. 26).

Figure 5 depicts the ramifications for moral agency. As situations become more complex, agents have to master higher level morality to cope successfully with these situations, and they have to manage to establish and uphold moral games by employing the respective positive (and sometimes negative) sanctions. In particular, they also have to be prepared to revert to lower forms of morality (1) in order not to victimize themselves and (2) in order to impose constraints on the other to bring them back to the path of virtue in terms of the higher stages.

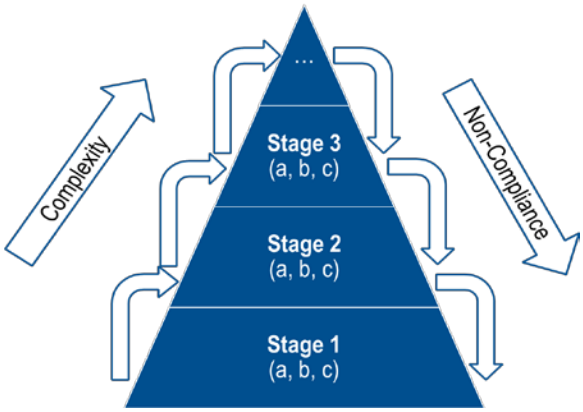

Figure 3: The dynamics of stage usage depending on complexity of the situation and compliance of the other agents

Altogether, I have mentioned six moral stages and the respective principles that form two triads, one grounded on sympathy, the other on conflicting interests. In the context of sympathy, the sanctions are issued in terms of *liking* and *dislike*. In the context of conflicting interests, the sanctions are issued in terms of *respect* and *disrespect*. Apart from these sanctions that relate directly to the content of the specific moral principle, an extended form of punishment is possible. It applies in situations, in which the other is either unwilling or unable to "play by the rules" of this game. In such a case, one can always punish by moving one stage down to play the lower level game. For instance, if someone fails to share with me, I would stop sharing at some point, which means that I move down to the game in which only the simple property rule applies and where everybody stays with what they have.

6. Ethics and economics – a taxonomy for the social sciences

If my analysis is correct and moral principles function as institutions, we end up with an economic theory of morality that not only shows how morality is implemented in moral agency, but it also internalises morality as such (i.e. as solutions concepts for mixed-motive games). However, does this mean that ethics is absorbed by economics? The answer is: yes and no. Yes, mixed-motive games constitute moral problems, and moral principles as solution concepts for these games are, in fact, institutions in the institutional economic sense. However, there is yet another perspective on ethics that is not captured by this reconstruction. This is the quest for justice.

Whether or not a certain state (or consequence of some act) is just in the moral sense is independent from the question of whether and how that state can be achieved. Here we have to be careful to distinguish different meanings of morality. For instance, when Turiel defines morality in terms of prescriptive judgments "about welfare, justice, and rights … that involve concern with dignity, worth, freedom, and treatment of persons" (2006, p. 10), this notion of morality pertains to the values a moral person has internalized. When Haidt takes morality as "interlocking sets of values, virtues, norms, practices, identities, institutions, technologies, and evolved psychological mechanisms that work together to suppress or regulate self-interest and make cooperative societies possible" (Haidt, 2012, p. 270), this seems rather vague, but stresses much more the aspect that morality consists in rules, i.e. institutions. Whereas the former concepts relates to decision-theoretic view, the latter relates to the game-theoretic view as discussed above.

There are clearly (at least) these two distinct meanings of morality, and the difference seems vitally important. *Moral principles as institutions* are tools to solve social dilemmas. They belong to the category of strategic reasoning from an objective, trans-individual point of view, i.e. transcending the inter-individual perspective of players involved in a game a taking the perspective of the game from the point of view of some choreographer. The players have to be able to take on this view to establish an institution (and if currently involved in the respective game, they have to leave it for a moment to think and reshape it). However, apart from requiring such move to a trans-individual perspective, this remains within the confines of strategic or technological thinking and poses a purely economic problem (i.e. one of maximizing utility). I call this strategic kind of reasoning "prescriptive" to distinguish it from normative questions in the ethical sense.

As opposed to this, *moral (or ethical) principles as ideas of justice* are truly "normative". It is common in economics to differentiate between positive and normative theory, but as it appears we have to distinguish between positive (or explanatory), prescriptive (or technological) and normative (or ethical) theory. Normative theory is about answers to questions concerning the "good life". Table 2 reveals that such questions can be asked with respect to a single individual, with respect to a group of individuals and with respect to everybody (i.e. including the perspective of every individual affected by a certain decision or state of affairs).

Table 3: A taxonomy for the social sciences (in particular economics and ethics) presents a systematic differentiation of scientific questions in two dimensions. The first dimension pertains to regulative principles, i.e. "truth" with respect to explanatory questions (positive theory), "efficiency" with respect to strategic questions of optimizing the outcome of actions (prescriptive theory), and the "good life" with respect to what we should aim at or what preferences we ought to have (normative theory).

Table 3: A taxonomy for the social sciences (in particular economics and ethics)[6]

Theoretical Perspective (Regulative Principle):	Explanatory (Truth)	Technological (Efficiency)	Ethical (Good life)
Social Perspective:			
Intra-individual (Decision Theory)	Positive DT (Behav. econ.)	Prescriptive DT (Rational choice)	Normative DT (Values, virtues)
Inter-individual (Game theory)	Positive GT (Behav. econ.)	Prescriptive GT (Non-coop. GT)	Normative GT (Cooperative GT)[7]
Trans-individual (Soc. systems theory)	Positive SST (Macro econ.)	Prescriptive SST (Institutional econ.)[8]	Normative SST (Theories of justice)

The second dimension pertains to the social level to which one refers. The intra-individual perspective asks how we have to explain an individual's behavior (positive), what an individual with certain preferences should do under specific restrictions (prescriptive) or what preferences an individual should have in terms of what accounts for a good life (normative). The inter-individual perspective concerns the question of how to explain the players' choices in games (positive), what strategies are rationalizable in games in terms of Nash-equilibria (prescriptive), and what is the best outcome for coalitions of players

[6] The table does not include the different fields of psychology and sociology that would also have to fit into this schema. However, it could be easily extended in this way. Here the main concern was to reveal how economic subdisciplines and in particular the field(s) of ethics relate to each other.

[7] Cooperative game theory belongs to this cell only insofar its solution concepts are concerned, which determine optimal states for coalitions of players.

[8] This would include, of course, morality in terms of institutions as explained above, but also theories of mechanism design in the social context.

in terms of cooperative game theory (prescriptive). Finally, the trans-individual perspective pertains to the analysis of social systems and how they function (positive), institutions in terms of what is to be done on the social level in order to solve problems cooperation problems that arise in non-cooperative game theory (prescriptive), and to what is just from an objective point of view (normative). The latter is the realm of ethics as we understand it today, and this is clearly independent from the various fields of economics that are included in this taxonomy.

In traditional philosophy of science, especially Popper's "logic of scientific discovery" (1959), positive theory was considered to be the only "true" science. However, more recent developments have revealed two important problems: On the one hand, positive theory is not entirely value-free, because we cannot determine how to assign truth-values to theories or statements based on deduction alone (see e.g. Putnam, 2002; see also Minnameier, 2017). And on the other hand, it has become clear that prescriptive and normative questions are questions in their own right, so that e.g. engineering sciences are not merely applied natural science just as the different interpretations of rational choice theory (typically labelled "positive" and "normative") relate to different research questions, of which each has its own dignity (Putnam, 2015). The presented taxonomy tries to capture all this and at the same time carve out systematic fields of research both within economics and beyond.

7. Ramifications and conclusion

Based on what has been developed in this paper, we not only attain a framework for systematic disciplinary and inter-disciplinary research, but we can also solve a few riddles that seem to have haunted us. Some have been mentioned in this paper, and I would like to address them in this final section.

For instance, we wonder why people generally imbued with some sense of morality, are susceptible to moral hypocrisy to a great extent. The puzzle obviously derives from a confusion of prescriptive and normative questions. From a normative point of view it is (perhaps) clear that throwing a dice is a fair way to determine who should be assigned a favorable or unfavorable task. However, for this to become a rule of the game in terms of game theory, it would have to be implemented as an institution, which it cannot, because the sanctions cannot be played. By the same token, if it is not (sufficiently) transparent whether and agent has made a particular choice or whether it is the result of a machine intervention as in the "plausible deniability conditions" employed in Dana, Weber and Kuang (2007) or Andreoni and Bernheim (2009), a certain morality does not become a rule of the game, and hence players play according to the actual rules that hold. That's all.

However, one may also ask, why a significant portion of people cooperate in the PD, even after sufficient experience with the game. Quite obviously, these individuals have strong preferences to cooperate, which makes it rational for them to cooperate even at the cost of permanently losing out to more self-interested players. Here, we might discuss (from a normative point of view), whether people ought to have these preferences, but as long as they have them, we can not only explain their behavior in the positive sense, but it seems even rationalizable in the prescriptive sense, given these preferences that are often referred to as part of a "moral identity" (e.g. Aquino & Reed, 2002; Walker, 2014). Actually, it seems questionable, indeed, to me from a normative point of view, but when we argue in a prescriptive context, preferences are "imported" as exogenous drivers and not to be questions in this frame of reference.

To sum up, differentiating the different fields of research opens up many fruitful routes to address systematically different questions, explain many previously surprising facts, integrating much of ethics into economics while at the same time carving out the proper realm of ethics and its relation to economics.

References

Andreoni J., & Bernheim D. B. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica, 77,* 1607–1636.

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology, 83 (6),* 1423–1440.
Aumann, R. J. (1974). Subjectivity and correlation in randomizing strategies. *Journal of Mathematical Economics, 1 (1),* 67-96.
Aumann, R. J. (1987). Correlated equilibrium and an expression of Bayesian rationality. *Econometrica, 55 (1),* 1-18.
Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics, 11,* 122-133.
Beck, K. (2008). Moral judgment in economic situations – Towards systemic ethics. In F. Oser & W. Veugelers (Eds.), *Getting involved: Global citizenship development and sources of moral values* (pp. 359-370). Rotterdam: Sense.
Becker, G. S. (1976). *The economic approach to human behavior*. Chicago, IL: University of Chicago Press.
Berg, J., Dickhaut, J., & McCabe, K. (1994). Trust, reciprocity, and social history. *Games and Economic Behavior, 10,* 122-142.
Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge, UK: Cambridge University Press.
Bicchieri, C. (2017). *Norms in the wild: How to diagnose, measure, and change social norms.* New York: Oxford University Press.
Binmore, K. (2010a). Social norms or social preferences. *Mind & Society, 9,* 137-159.
Binmore, K. (2010b). Game theory and institutions. *Journal of Comparative Economics, 38,* 245-252.
Binmore, K. (2011). *Natural justice*. Oxford: Oxford University Press.
Brennan, G., & Buchanan, J. M. (1985). *The reason of rules: Constitutional political economy*. Cambridge, MA: Cambridge University Press.
Bruni, L., & Sugden, R. (2007). The road not taken: How psychology was removed from economics, and how it might be brought back. *The Economic Journal, 117 (1),* 146-173.
Cappelletti, D., Goth, W., & Ploner, M. (2011). Being of two minds: Ultimatum offers under cognitive constraints. *Journal of Economic Psychology, 32,* 940–950.
Colby, A., & Kohlberg, L. (1981). *The measurement of moral judgment, Vol. 1: Theoretical foundations and research validation.* Cambridge, MA: Cambridge University Press.
Cornelissen, G., Dewitte, S., & Warlop, L. (2011). Are social value orientations expressed automatically? Decision making in the dictator game. *Personality and Social Psychology Bulletin, 37,* 1080–1090.
Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic theory, 33,* 67-80.
Darwall, S. (2005). *The second-person standpoint: Morality, respect, and accountability*, Cambridge, MA: Harvard University Press.
Deci, E. L., & Ryan, R. M. (2000). *Self-Determination theory: Basic psychological needs in motivation, development and wellness.* New York: The Guilford Press.
Deci, E. L., & Ryan, R. M. (2000). The „what" and „why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11 (4),* 227–268.
Dhami, S. (2016). *The foundations of behavioural economic analysis*. Oxford: Oxford University Press.
Dietrich, F., & List, C. (2013a). A reason-based theory of rational choice, *Noûs, 47 (1),* 104-134.
Dietrich, F., & List, C. (2013b). Where do preferences come from? *International Journal of Game Theory, 42,* 613-637.
Ellingsen, T., Johannesson, M., Mollerstrom, J., & Munkhammar, S. (2012). Social framing effects: Preferences or beliefs? *Games and Economic Behavior, 76 (1),* 117-130.
Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism: Experimental evidence and new theo- ries. In S. Kolm & J. Ythier (Eds), *Handbook on the economics of giving, reciprocity, and altruism, Vol. 1* (pp. 615-669), Amster- dam: Elsevier.
Frankena, W. (1980). *Thinking about morality*. Ann Arbor: University of Michigan Press.
Fudenberg, D, & Tirole, J. (1991). *Game theory*. Cambridge, MA: The MIT Press.
Gert, B. (2005). *Morality: Its nature and justification*. Revised ed., New York: Oxford University Press.
Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them.* New York: Penguin.
Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. London: Allen Lane.

Haidt, Jonathan. 2001. The emotional dog and its rational tail. Psychological Review 108: 814–834.
Hare, R. M. (1981). *Moral Thinking*. New York: Oxford University Press.
Hobbes, T. (1651/2001). *Leviathan.* South Bend, IN: Infomotions.
Hoffman, E., McCabe, K., & Smith, V. (1996). Social distance and other-regarding behavior. *American Economic Review, 86,* 653-660.
Homann, K. (2016) Theory strategies of business ethics. In C. Luetge & N. Mukerji (eds.), *Order ethics: An ethical framework for the social market economy* (pp. 37-54), Cham: Springer International Publishing Switzerland.
Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology, 32,* 865-889.
Kahneman, D. (2011). *Thinking, fast and slow*. London: Penguin.
Kohlberg, L. (1981*). Essays on moral development, Vol. 1: The philosophy of moral development*. New York: Harper Row.
Krebs, D. L., & Denton, K. (2005). Toward a more pragmatic approach to morality: A critical evaluation of Kohlberg's model. *Psychological Review, 112,* 629-649.
Ledyard, J. O. (1995). Public goods: A survey of experimental research. In J. H. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics* (pp. 111-194). Princeton, NJ: Princeton University
List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy, 115,* 482-494.
Luetge, C., Armbrüster, T., & Müller, J. (2016). Order ethics: Bridging the gap between contractarianism and business ethics. *Journal of Business Ethics, 136*, 687-697.
Minnameier, G. (2000). *Strukturgenese moralischen Denkens - Eine Rekonstruktion der Piagetschen Entwicklungslogik und ihre moraltheoretischen Folgen***.** Münster: Waxmann.
Minnameier, G. (2017). Forms of abduction and an inferential taxonomy. In L. Magnani & T. Bertolotti (Eds.), *Springer handbook of model-based reasoning* (pp. 175-195), Berlin: Springer.
Minnameier, G. (forthcoming). How to explain the 'Happy-Victimizer' in adulthood. *Frontline Research.*
Paulus, M., & Moore, C. (2014) The development of recipient-dependent sharing behaviour and sharing expectations in preschool children. *Developmental Psychology, 50 (3),* pp. 914-921.
Paulus, M., & Moore, C. (2017). Preschoolers' generosity increases with understanding of the affective benefits of sharing. *Developmental Science, 20 (3),* e12417.
Popper, K.R.: 1959, *The Logic of Scientific Discovery*, Hutchinson, London.
Putnam, H. (2002): *The collapse of the fact/value dichotomy and other essays*. Cambridge, MA: Harvard University Press.
Putnam, H. (2015). Naturalism, realism, and normativity. *Journal of the American Philosophical Association, 1 (2),* 312-328.
Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review, 118,* 57-75.
Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489 (7416),* 427-430.
Samuelson, P. A. (1948). Consumption theory in terms of revealed preference. *Economica, 15 (60),* 243-253.
Scanlon, T. M. (2014). *Being realistic about reasons*. Oxford: Oxford University Press.
Schelling, T. C. (1960). *The strategy of conflict.* London, England: Oxford University Press.
Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford library of psychology. The Oxford handbook of thinking and reasoning* (pp. 433-455). New York, NY, US: Oxford University Press.
Turiel, E. (2006). Thought, emotions, and social interactional processes of moral development. In M. Killen & J. Smetana (Eds.), *Handbook of moral development* (pp. 7-36). Mahwah, NJ: Lawrence Erlbaum Associates.
Walker, L. J. (2014). Moral personality, motivation, and identity. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 497–519). New York: Psychology Press.