On the ethics of public nudging: Autonomy and Agency

Preliminary Draft, please do not quote!

Christian Schubert
Chair of Economics & Ethics
University of Kassel
Nora-Platiel-Strasse 4
34109 Kassel, Germany
phone: +49-561-804 3857
dr.c.schubert@gmail.com

Abstract
Nudges, i.e., low-cost interventions that steer people's behavior without compromising their freedom of choice, are the key contribution of 'Libertarian Paternalism' (LP) to public policy. They typically work through either harnessing or responding to people's cognitive biases and heuristics – which is why they have been criticized for being manipulative and for compromising personal autonomy. We argue, though, that (i) nudging hardly compromises autonomy, properly understood, and that (ii) it rather risks undermining people's *agency*, i.e., their ability to engage in creative self-constitution over time. This reorientation has far-ranging implications for the ethics of behavioral policies in general and LP in particular.

> ”In this century, it seems to me, our greatest enemy will not be drones or ISIS or perhaps even climate change: It will be convenience.” (Andrew Smith)

1. INTRODUCTION

Ever since its introduction, in 2008, by Cass Sunstein and Richard Thaler, the concept of ’nudging‘ has stirred a lot of discussion in academia, policy circles, and the general public.[1] The term stands for policy interventions that aim at influencing people’s behavior without changing their choice sets. Through the design of the *choice architecture* (i.e., their decision context), agents are supposed to be steered in a particular direction, while retaining the freedom to choose otherwise: To illustrate, a cafeteria manager can influence his clients’ choices through the way he arranges the food on display; an employer can foster her employees’ savings decisions by making enrolment in 401(k) pension schemes the default, with an easily available ‘escape clause’.[2] Nudges can however also be used for non-paternalistic purposes, such as promoting pro-environmental behavior (e.g. Pichert and Katsikopoulos 2008, Schubert 2016). Since they typically work their magic by purposefully harnessing people’s cognitive biases, many critics find nudges objectionable for treating people disrespectfully by compromising their personal *autonomy*. We can expect Big Data to widen the realm of applications considerably in the future, thereby giving rise to interesting ethical questions.

This paper discusses the ethics of public (i.e. government-issued) nudging, in particular its normative costs in terms of autonomy or self-government on the one hand and *agency* on the other hand.[3] Specifically, we ask two questions:

(i) Do nudges compromise the autonomy, properly understood, of the agents exposed to their influence, or rather some other value?

---

[1] See in particular Sunstein and Thaler (2003), Thaler and Sunstein (2003, 2008), Sunstein (2014). See Rebonato (2012: 257f., endnote 1) on the substantial impact that the overarching program of ‘Libertarian Paternalism’ has had on public policy in the U.S., the U.K. and elsewhere. As to the U.S., see the executive order No. 13,707, 80 Fed.Reg. 56, 365-66 (Sept. 15, 2015), http://perma.cc/FDR2-VX3T. Related approaches to behavioral policymaking are Camerer et al. (2003) and Loewenstein and Haisley (2008). For what it’s worth, preliminary evidence shows that nudges may be more popular than most critics care to concede (e.g., Hagman et al. 2015; Tannenbaum et al. 2014).

[2] See Thaler and Sunstein (2008: 1-3), Thaler and Benartzi (2013).

[3] A note on terminology: In what follows, ‘nudges’ will be understood as public nudges; ‘autonomy’ will be understood in the sense of accountability-conferring ‘personal autonomy’ (as opposed to ‘legal autonomy’, say) and as a *capacity*, i.e. as a matter of degree (as opposed to, say, an absolute right, see Feinberg 1986). And by ‘agency’, we refer to the individual’s capacity to constitute herself as an agent, a task that takes a lifetime (Korsgaard 2009). Note that Sunstein (2015c) seems to use that term in a different way.

(ii) Which implications follow from the answer to (i) with respect to the design of people's choice architecture (of which nudges are a part)?

There is a growing interest, not only in the intricacies of the *welfare* effects of nudges (e.g. Grüne-Yanoff 2010), but also in the impact of nudging on autonomy. A seminal paper is Bovens (2009); other key contributions include Hansen and Jespersen (2013), Hausman and Welch (2010), Rebonato (2012), Schnellenbach (2012, 2016), and Wight (2013).[4] Advocates of nudging have only recently joined the debate on the potential normative costs of nudging (Sunstein 2015a, Sunstein 2015c).

This debate, however, suffers from several shortcomings, two of which stand out.[5] First, nudges as such (i.e., as an instrumental method) are often confounded with *Libertarian Paternalism* (henceforth LP), i.e., the overarching normative policy program in which Thaler and Sunstein have embedded them – typically, when critics attack nudges, what they really seem to target is LP, in particular its paternalistic ('nanny state') aspects.[6] In a nutshell, LP says that nudges should be used to 'improve' agents' choices – specifically, to steer agents in the direction of those choices they would have made were they perfectly rational and 'fully informed'; then, it's only a small step towards deeming paternalistic interventions legitimate.[7] As a consequence of the confusion of nudges with LP, the specific ethical issues associated with nudges *per se* are often neglected. Second, most advocates and critics of nudges ground their arguments on conceptions of autonomy that are not coherently applicable in a 'behavioral world', i.e., a setting where agents have limited mental resources (i.e., limited cognitive capacities, attention, and willpower) and incomplete and context-dependent – hence, often inconsistent – preferences. A behavioral world is a world where nudges actually work, and where any discussion of the ethics of nudging should be situated. As a consequence of these two problems, we still lack the basis for a proper weighing of the costs and benefits involved in nudging.

Both shortcomings are related: The ethics of nudging can best be discussed outside the narrow focus on LP that characterizes the debate so far. The key problem is that LP itself is a

[4] See also Barton and Grüne-Yanoff (2015), Binder (2014), Binder and Lades (2015), Fischer and Lotz (2014), Hansen (2015), Mills (2013, 2015), Nagatsu (2015), Selinger and Whyte (2011), and Smith et al. (2013).

[5] Another flaw is the widespread claim that a 'pretense of knowledge' is necessarily involved in public nudging. As we argue in the following, there may be a variety of life domains where individuals no longer enjoy 'epistemic privilege' regarding their own preferences: Government may (in the near future anyway) have access to technologically advanced behavioral algorithms. Arguably more disturbing is the fact that policymakers and regulators are subject to the same *cognitive biases* as everyone else (see Schnellenbach and Schubert 2015).

[6] This can best be seen by consulting the critics' stance toward non-paternalistic or 'social' nudging, which tends to differ markedly from their evaluation of paternalistic nudging (e.g. Wight 2013: 110f.).

[7] As Selinger and Whyte (2011: 924) put it, LP wants to 'channel' people's biases 'and put their influence to good use.'

welfarist, strictly outcome-oriented program that assumes given preferences[8] – marred by given biases –, thereby disregarding the peculiarities of processes of preference formation. Autonomy and agency, though, are all about the formation of preferences: Strictly speaking, they don't concern the nominal freedom to act upon one's given preferences (that's *freedom of choice*),[9] but rather the way an agent forms them in the first place and follows through on them – *freedom of will* (Fischer and Lotz 2014), hence something economics is typically silent about: For instance, economists' somewhat truncated understanding of autonomy as 'consumer sovereignty' simply posits to respect an agent's given, complete and consistent preferences as the only source and ultimate measuring rod of normative judgments.[10] Quite obviously, this is of little help in a behavioral world (Sugden 2004).

As we will show, a closer look at the issue of human preference formation with limited mental resources reveals not only that human preferences are context-dependent (in behavioral economics parlance: they are *constructed*), but also that real-world individuals *need* to rely on contextual features – of an institutional, situational and interpersonal kind – in order to either construct their preferences in the first place, to make them effective (i.e., to act upon them), and ultimately to constitute themselves as agents. In other words, agents exhibit, as a key consequence of their bounded rationality, what Davis (2014) refers to as 'bounded individuality': Their preferences – and, hence, their identity – are partly endogenous to their environment.[11] In light of this, we submit that what's really at stake in nudging is people's *agency*, which we define – following Korsgaard (2009) – as people's capacity to engage in the ongoing, specifically human, project of identity formation or self-constitution. In terms that might sound paradoxical, nudges may give rise to 'excessive convenience'.

The argument proceeds as follows. Section 2 prepares the conceptual ground for discussing the impact of nudging on autonomy. Section 3 gives a critical overview of autonomy conceptions used in the extant literature on the ethics of nudging. We suggest to overcome the issues identified there by introducing an alternative account of autonomy suggested by Sarah Buss. In section 4, drawing on work by James Buchanan and Christine Korsgaard, we submit that what's really at stake in the ethics of nudging are its effects on people's agency. We also suggest some implications. Section 5 concludes.

---

[8] Let's save words and understand 'preferences', henceforth, as also encompassing values.

[9] In order to clarify the ethical debate on nudging, it's crucial to keep these values apart. With respect to the freedom dimension, most contributions to the debate presuppose negative freedom. Discussing nudging from the viewpoint of alternative notions of freedom obviously deserves further research (but see Grüne-Yanoff 2012, Mills 2013).

[10] See, e.g., the principle of 'preference autonomy' suggested by Harsanyi (1982: 55): 'In deciding what is good and what is bad for a given individual, the ultimate criterion can only be his own ... preferences.'

[11] See also FN 49, below.

2. PREPARING THE GROUND

Before we can discuss the conceptions of autonomy used in the critical literature on nudging (we'll do that in section 3), some conceptual groundwork is in order. Two things, in particular, need to be clarified at the outset.

First, what are nudges anyway? There has been some controversy on how to specify this innovative set of public policy instruments. A key problem is that nudges are often defined in a very broad way, to also encompass the mere provision or disclosure of factual information, a measure that not only lacks originality (Rebonato 2012), but that also does not raise interesting ethical questions. To illustrate, Thaler and Sunstein (2008: 6) define nudges as 'any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a nudge, the intervention must be easy and cheap to avoid.' This definition has been criticized extensively.[12] We submit to follow Hansen (2015) and focus on the notion, originally advanced by Thaler and Sunstein (2008: 8, emphasis added) themselves, that 'a nudge is any factor that significantly alters the behavior of Humans, even though *it would be ignored by Econs*', where 'Humans' refers to real-world individuals with limited mental resources, and 'Econs' refers, basically, to homo oeconomicus. Adding the crucial aspect of intentionality, we suggest defining nudges as interventions that aim at altering people's behavior by either (i) harnessing cognitive biases or (ii) responding to them. To illustrate, consider the deliberate change of defaults on the one hand, and the introduction of cooling-off periods on the other hand. Understood this way, nudges should be considered the actual original and imaginative contribution of LP to public policy-making.

Crucially, all definitions offered so far include a basic *transparency* requirement – Thaler and Sunstein (2008: 244), for instance, suggest the Rawlsian 'publicity principle', according to which government should be banned 'from selecting a policy that it would not be willing or able to defend publicly to its own citizens.'[13] While this particular way to

[12] See, e.g., Hausman and Welch (2010), Hansen (2015), Hansen and Jespersen (2013), and Mongin and Cozic (2014) on the way in which Thaler and Sunstein define and use several key terms. For instance, the aspect of intentionality is missing in the first sentence quoted above – it's, however, crucial to distinguish mere choice architecture from purposeful nudging (Barton and Grüne-Yanoff 2015). Note that Sunstein himself still sometimes confounds both, as e.g. in Sunstein (2015a: 6, 9). Barton and Grüne-Yanoff (2015) define a nudge as 'an intervention on the choice architecture that is predictably behaviour-steering, but preserves the choice-set and is (at least) substantially noncontrolling [i.e., there's an easily available escape clause, C.S.], and does not significantly change the economic incentives', which is still somewhat unsatisfying ('economic'?).

[13] See Hansen and Jespersen (2013: 23-27) for a critical discussion on this approach.

incorporate transparency in nudging is certainly unsatisfactory, let's stipulate for the sake of the argument that *genuine* (public) nudging comes with what Bovens (2009: 217) refers to as 'in principle token transparency'.[14] In other words, the escape clause must not merely be nominally present, but effectively available, if possibly at some cognitive cost. To be sure, transparency makes most nudges somewhat less effective. That effect should however not be overestimated (Loewenstein et al. 2014) – a point often neglected by the critics (e.g. Wight 2013: 134f.).

Our definition lets the purposeful design of defaults, graphic images on billboards or cigarette boxes, manipulated road markings that make motorists slow down – near dangerous curves, say –, social advertising campaigns that invoke social norms, 'prompted', i.e., voluntary, choice, reminders, warnings, and brief cooling-off periods – all subject to transparency requirements (however specified in detail) – qualify as nudges. On the other hand, measures such as disclosure requirements that provide individuals with information in a perfectly neutral fashion – consider calorie counts on restaurant menus or Sunstein's own pet example, GPS –, as well as measures that try to rationally *persuade* agents to choose certain actions (Hausman and Welch 2010) don't make it on the list. It's an open question whether nudges should be distinguished from what Grüne-Yanoff and Hertwig (2015) refer to as 'boosts', i.e., interventions that try to help people improve their decision-making (rather than their choices) by overcoming (rather than harnessing or responding to) their cognitive biases.[15] Anyhow, both nudges and boosts, but also mandatory choice (Rebonato 2012: 134-141), and ultimately all measures that involve the (re-)design of people's choice architecture on the basis of psychological insights, may be subsumed under the general category of *behavioral policies* or *behavioral interventions*.[16]

Second, in a behavioral world, agents require some nudging. To see why, consider the way real-world individuals go about constructing their preferences. Given the fact that they have limited and easily exhausted mental resources, they cannot engage in preference construction without contextual support.[17] In other words, every human agent needs to rely, in one way or another, on her surrounding choice architecture in order to economize on scarce

[14] This principle requires that a hypothetical 'watchful' individual should be able to identify the nudge as well as the underlying intention and 'could blow the whistle if she judges that the government is overstepping its mandate' (ibid.). This arguably excludes tools such as subliminal advertising. As Anderson (2010: 374) points out, Thaler and Sunstein in general do not seem to take transparency requirements very seriously. See Sunstein (2015a: 19, 38) on that issue.

[15] To be sure, this is controversial as well; for starters, Sunstein (2015c) doesn't see an opposition between nudges and what he refers to as 'education'.

[16] To illustrate, the 'new policy tools' described by Chetty (2015: 6-12) are examples of behavioral policies.

[17] Consider the subfield of behavioral economics studying 'ego depletion' (e.g. Baumeister and Tice 1998).

mental resources; this necessarily includes some (intentional) nudging, especially so when confronted with complex or new decision problems (Schubert 2014). In the modern world in which we live, a significant part of the decision context has been deliberately created by choice architects self-conscious about their role. In the words of Sunstein (2014), '[e]very hour of every day, choices are implicitly made for us, by both private and public institutions, and we are both better off and more autonomous as a result. If we had to make all decisions that are relevant to us, without the assistance of helpful choice architecture, we would be far less free ... choice architecture enables us to be free … If we had to make far more decisions, our autonomy would be badly compromised, because we would be unable to focus on what concerns us.' (ibid.: 130f., 137).[18]

In order to link nudges to autonomy (and, eventually, to agency), it may be useful to redescribe what is involved in nudging: Let's borrow a term introduced by Valdman (2010) and understand being nudged as the partial *outsourcing of self-government* to an agent's context.[19] Arguably, *all* nudges involve some kind of partial outsourcing by making life's chores convenient enough to cope with.[20] Given this understanding, when assessing a particular nudge, we have to ask the question: Can we reconstruct that nudge as the product of an act of voluntary partial 'outsourcing of self-government' to some external body?[21] To the extent that nudges can plausibly be so reconstructed, two remarkable things happen: First, the gap between their paternalistic and their non-paternalistic use closes;[22] second, it becomes conceivable that agents consent to 'manipulation', in the sense of letting their own cognitive biases be deliberately harnessed.[23] If that's the case, nudges do not treat people as mere means, *pace* Wight (2013: 135). We might instead say that under these – notably ideal – circumstances, nudges can be reconstructed as tools of 'behavioral self-commitment' or 'self-

[18] See also de Marneffe (2006: 81). All this flies in the face of the standard economist's assumption that autonomy is positively correlated with the size of an agent's opportunity set (i.e., the accuracy of perceived opportunity costs).

[19] In Valdman (2010: 762), the (total) outsourcing of self-government means 'relinquishing your final authority as the arbiter of your actions', not just 'deferring to some advisor's judgment'. You may read this as a - dystopian - case of technologically perfect nudging.

[20] To see this, consider the harnessing of framing effects: To the extent that the agent lets himself be influenced by them, he 'delegates' parts of his deliberation to some contextual factor (i.e., a part of the choice architecture). Analogous reasoning applies to the impact of graphic images, but also of prompted choice: in these cases, the effort to overcome self-control issues is partly delegated. With deceiving road markings, the effort is also delegated in the sense that the choice architecture relieves the motorist from cognitive overload. In the case of cooling-off periods, the agent can be described as delegating parts of his deliberation to his own, future, self.

[21] That question is rarely asked in the literature - but see Kirchgässner (2015: sect. 3) and Schubert (2014).

[22] Paternalism is standardly defined as a welfare-improving intervention that's *unwanted* (e.g. Dworkin 2014).

[23] 'Manipulation' is a value-laden term. Let's define it, crudely, as the deliberate harnessing of an agent's biases using stealth. Genuine (i.e. transparent) nudging can only be manipulative in the limited sense that, say, being the voluntary object of emotional seduction might be characterized as being manipulated (Buss 2005: 212f.); As Sunstein (2015b: 7f.) puts it, 'within limits, being manipulated can even be fun. In some forms, manipulation is a form of play.' (see also Sugden 2008: 247). See also Wilkinson (2013: 353f.) on consensual manipulation.

nudging' (Lades 2014) that people deliberately employ to further their own goals. Consider re-arranging the food in your fridge in order to fight unwanted temptation. To be sure, doing so requires skills, perhaps particularly the ability of self-control (which may, of course, be supported in turn by a successful self-nudge).

So, human agents always require some nudging. This implies that when debating the ethics of nudging, we should stop idealizing the institutional status quo. Most critics seem to implicitly assume that before the implementation of public nudges, people act upon preferences that are somehow 'authentic', and that public nudging then spoils the show by distorting processes of preferences formation. This is misleading. For in a behavioral world, processes of individual preference formation are highly context-dependent, i.e., they are heavily influenced by the given choice architecture, and in a mostly subconscious manner to boot (e.g. Felsen and Reiner 2015). As Reiss (2013: 299) puts it, 'humans with bounded rationality and willpower are subject to myriad influences anyway, and most of them do not aim to improve consumer well-being' – which implies that the notion of 'authentic' preferences does not make conceptual sense (Fischer and Lotz 2014: 11).[24] The institutional status quo is *not* characterized by a perfectly 'neutral' choice architecture, but rather by set of contextual factors that, while partly the result of random (thus, 'neutral') processes, are also the product of intentional and often harmful *private nudging* – most likely promoted rather than discouraged by market forces (Akerlof and Shiller 2015, Bar-Gil 2012)..[25] Consequently, nudging should be understood as a marginal *redesign*, *adjustment* or *correction* within a specific domain's given, highly complex choice architecture, rather than as the introduction of an additional, hitherto unknown kind of 'distortion'.[26]

Our two clarifications – nudging (properly understood) answers to cognitive biases, and there are no 'authentic' preferences in the first place – allow us to finally take a critical look at the autonomy conceptions that seem to inform the extant ethical debate on public nudging.

[24] Talking about 'undistorted' preferences might also suggest that only 'natural' or 'non-artificial' preferences (whatever that means) deserve to be respected, a view rightly criticized by Hayek (1961).

[25] Note that while the non-neutrality of the given choice architecture is certainly unavoidable, *nudging* – understood as the deliberate attempt to steer people in certain directions without curtailing their choice sets – is not. It's true that, say, a default is sometimes inevitable, but that does not imply that it must be chosen with the aim of steering people in certain directions. Moreover, it may often be possible to replace it by mandatory choice.

[26] The fact that the redesign of a given choice architecture is marginal does of course not imply that the behavioral effect induced by it is only 'marginal'.

3. 'AUTONOMY' IN THE DEBATE ON NUDGES

Autonomy is, of course, a key component of most conceptions of the good life – a part of what makes a life worth living[27] – and widely perceived to be the key ethical value at stake in nudging. In order to probe whether a given conception of autonomy can be coherently applied in the context of nudging, we will examine whether it can be coherently applied in our behavioral world. Let's call this standard the criterion of external coherence (EC). Most contributions to the ethics of nudging seem to reflect what may be referred to as the received view on autonomy.

3.1 How autonomy appears in the ethics of nudge literature

What exactly does it mean to enjoy self-government, to determine the direction of one's own life, to 'own' one's actions in a way that grounds accountability? What's the special self-relation that makes an agent act 'autonomously'? Apparently, some of the authors mentioned above share the view that a given agent acts autonomously to the extent that she's a 'super-agent' (Buss 2012: 656, 678): She puts a lot of critical reflection into the formation and adjustment of her motives, which makes her respond the right way to changes in incentives. Let me give two examples.

Hausman and Welch (2010: 128, FN 16) take autonomy to mean 'the *control* an individual has over her own evaluation, deliberation and choice' (my emphasis; see also Mills 2013: 30-32). Nudges would then compromise an agent's autonomy to the extent that they influence her behavior by harnessing certain cognitive biases, rather than by trying to engage her deliberative faculties (through means of rational persuasion, say). In other words, by circumventing the agent's deliberative faculties, the control she exercises over her own will would be undermined (ibid.: 130), making her a passive bystander to her own actions.[28] To illustrate, on this view a purposeful change of defaults would qualify as autonomy-compromising to the extent that its power is due to the fact that defaults make agents believe to somehow 'possess' the default option, thereby exploiting their loss aversion (Smith et al.

[27] Important recent exceptions are Valdman (2010) and Conly (2014).

[28] See Wight (2013: 133-135) for a closely related view, implying that to the extent that qua the exploitation of cognitive biases, nudges lead to choices that are "not entirely" one's own (ibid.: 135). It's doubtful, though, whether the example he uses to illustrates that point – a rule restricting cigarette sales to one day per week – qualifies as a nudge in the first place.

2013). On the other hand, cooling-off periods may, on this view, enhance people's autonomy (Hausman and Welch 2010: 132f.).

Bovens (2009) resorts to a closely related conception of autonomy: According to him, an agent acts autonomously to the extent that her actions are responsive to *reasons*,[29] i.e., to principles that she can underwrite (ibid.: 209f.).[30] For instance, framing effects would, on this view, count as mere *causes* that reliably influence behavior, rather than as reasons. To illustrate, people tend to be more willing to donate organs if they are given information about the percentage of others who are already registered as donors, as opposed to information about those who are *not* registered (ibid.: 208f.). To the extent that nudges make use of such non-rational causal factors, then, they would compromise people's autonomy.[31] Bovens goes on to argue that the coherence (or rather lack thereof) of the preference structure an agent ends up with after having been exposed to nudging may serve as an indicator of autonomy losses. With fragmented preferences, an agent risks being eventually unable to recognize herself in her own actions (ibid.: 212-14). Aesop's famous fox is a case in point (Elster 1982): Seeing that some high-hanging grapes he originally longed for are out of reach, he dismisses them as 'sour anyway', thus displaying adaptive (read: incoherent) preferences, rather than genuine preference change – by acting 'out of character', he gains in terms of welfare, while losing in terms of autonomy.[32]

So far, the autonomy concepts used are familiar to moral philosophers. Yet, an alternative account has been suggested that goes beyond the received view by focusing on an individual's *subjectively perceived* autonomy: When people feel manipulated, they suffer in terms of procedural utility – for the subjective experience of self-determined choice can be hedonically valuable (e.g. Frey et al. 2004). Hence, a consistent welfarist would need to account for such concerns. There's also empirical evidence that nudges lose their effectiveness if agents feel manipulated in forming their own preferences (LeGrand and New 2015: 108-110, referring to Deci and Ryan 2000). LeGrand and New (2015: ch. 7) even argue

[29] In Bovens' words: when my action is reason-driven, then what drives my action is a feature of that action that I endorse as a feature that makes the action desirable (ibid.).

[30] That's the well-established conception of autonomy as 'responsiveness to reasoning' or 'responsiveness to reasons'. Stuart Hampshire (quoted in Buss 2012: FN 5) summarizes the former account: 'The more the sequence of a man's own ideas can be explained without reference to causes outside his own thinking, the more... self-determining he is.' See Buss (ibid.: 650-52) for a critical assessment. A somewhat 'economized' variant of the latter account is suggested by Binder and Lades (2015) who seem to equate autonomy with neoclassical utility maximization.

[31] See also Nagatsu (2015: sect. 4) on using responsiveness to reasons as a criterion for autonomous action.

[32] Another manifestation of the 'super-agent' view of autonomy, due to Frankfurt (1971), requires the agent to adopt some 'higher' standpoint and *endorse* (or identify with or at least to be non-alienated toward) not just the psychological elements involved in the formation of her preferences, but rather her preferences themselves or, alternatively, their historical genesis (e.g. Christman 2005). To our knowledge, this account has not yet been explicitly used in ethical discussions on nudging, despite its popularity in contemporary ethics (Buss 1994: 95f.).

that the debate on nudging should focus on its impact on people's *perceived* rather than their actual autonomy (see also Rebonato 2012: sect. 8.3).

Let's simplify things and denote the set of psychological factors that enter the process of preference formation as F. Then, the accounts of autonomy used so far in the literature on the ethics of nudging can be roughly summarized as defining autonomous action through a focus on (i) the power exercised by the agent's 'self' over F ('control'), (ii) the self's 'correct' way of processing *reasons* (i.e. a subset of F acknowledged as having that status), (iii) the quality of the product of F, viz., the preference or the action itself ('coherence'), and/or (iv) the subjectively perceived relationship between the agent's self and F ('procedural utility'). At first sight, these look like intuitive requirements of autonomous action. As we will argue, though, the corresponding accounts of autonomy cannot be coherently applied when it comes to assessing the normative costs of nudging. What, exactly, is wrong with them?

First, the notion of 'autonomy as control' presupposes a level of self-knowledge on the part of the agent that cannot be found in a behavioral world. A realistic model of man should take account of the fact that human behavior is typically influenced by a variety of causal factors operating at a subconscious level. This implies that the level of self-knowledge or self-transparency implicitly assumed in such accounts lacks descriptive accuracy. We rarely have access to the deep psychic sources of our motives (Buss 1994: 96).[33] The self can misunderstand itself. It may even *want* to misunderstand itself: Self-deception has been uncovered as an important source of well-being in behavioral economics.[34]

Second, the conception of 'autonomy as responsiveness to reasons or reasoning' violates the EC criterion by virtue of implicitly assuming that there is in fact a definitive set of 'correct' reasons or ways of reasoning that make the resulting preferences truly autonomous (Nagatsu 2015: sect. 4). Given the large variety of factors that enter processes of preference formation in a behavioral world, it seems arbitrary to single out some of them as 'autonomous', while dismissing all others as 'undue external influences'. In other words, this account is closely linked to specific – notoriously controversial – conceptions of rationality. To illustrate, the influence of framing is often seen as a paradigm case of heteronomy. This makes perfect sense if we accept homo oeconomicus as the role model of correct decision-making (as Thaler and Sunstein happen to do). However, things look differently from the

[33] As Christman (2008: 338) points out, 'only a marginal proportion of the self implicated in behavior and social interaction can ever be said to be available to conscious reflection … Hence, a person's inner picture of her motivational matrix can be highly incomplete and … inaccurate.' See also Felsen and Reiner (2015: sect. 3 and the references given therein). That's why accounts such as the one suggested by Watson (1987) cannot be applied in our context. According to him, autonomous action requires, inter alia, the capacity to 'reflect... on the origins of [one's] motivations' (ibid.: 152).

[34] See, for instance, the psychological insights on attribution bias (e.g. Heider 1958; Pal 2007).

viewpoint of alternative conceptions of rational behavior, such as the account of *ecological rationality* (Smith 2003, Berg 2014), according to which a choice that is inconsistent with the laws of logic can nonetheless qualify as 'ecologically rational' by virtue of reflecting a good enough fit between the agent's mind and her decision context. What looks 'irrational' from a standard economics viewpoint, may very well prove beneficial for the agent concerned (Gigerenzer 2015). Whichever concept of rationality we prefer – we should not build a conception of autonomy on the shaky grounds of controversial notions of rational action.[35]

Third, the notion of 'autonomy as coherence of the product of deliberation' can also be argued to violate EC: for instance, it's obviously not applicable when preferences are incomplete. Put differently, it seems bizarre to call someone heteronomous for that person's lack of complete preferences (see also Nagatsu 2015: sect. 4).

Fourth and finally, the introduction of subjectively experienced, procedural utility aspects quite obviously fails to offer us a reliable measuring rod that can be used to assess the normative costs of nudging. Subjective perceptions of manipulation can themselves be manipulated by a sufficiently sophisticated 'nudger' (such as the paternalistic expert committee imagined in Valdman 2010).[36] This account not only violates our EC criterion; it also points toward a violation of a complementary criterion, viz., *Independence of Epistemic Privilege*, which we will introduce presently (see section 4, below).

Hence, we have seen that the accounts of autonomy that are typically referred to in the extant literature on the ethics of nudging cannot be coherently applied in our context. Let's see whether there is an alternative account of autonomy that might do the job.

4. FROM AUTONOMY TO AGENCY

Sarah Buss has developed a conception of autonomous action that, we submit, overcomes the issues marring the accounts described above.[37] We will discuss it briefly in this subsection, concluding that when properly understood, autonomy can hardly be seen as the value that risks being compromised by nudging. Rather, what's arguably at stake is people's *agency*, as we will show in section 4.

[35] Buss (2012: 664f.) rejects this account of autonomy on yet other grounds: We can easily imagine cases where someone responds perfectly well to reasons, yet her resulting action can nonetheless hardly be judged autonomous (in the accountability-conferring sense) – to illustrate, consider the woman overcome by fear when threatened by a rapist.

[36] Consider the well-known experience machine described by Nozick (1974: 42-45).

[37] See in particular Buss (1994, 2005, 2012).

4.1 Against standard notions of autonomy

Buss rejects the 'super-agent' view of autonomy, sketched above, mainly for a simple reason: people are intuitively also held morally accountable for (most) actions they do habitually, akratically, unwillingly, or even thoughtlessly (Buss 2012: 651, 655).[38] It seems that neither *doing something* (such as: deliberating) nor *being satisfied with something* (the products of one's deliberation) can be considered the key to the particular self-relation that makes an agent act autonomously (ibid.: 656).

What's the alternative, then? Buss' key argument can easily be understood along the lines of our EC criterion: Since in a behavioral world, no agent can deliberately influence *all* her deliberations and endorsements, the factor that makes an agent act autonomously must be located in the passive ('nonagential') role he plays *when forming his preferences*.[39] Agents are inevitably passive in relation to most of the (disposition-shaping) psychological factors that causally influence their practical reasoning at any given moment.[40] Buss submits that someone's autonomy depends on the extent to which he 'can be identified with the direct, purely causal, nonrational influences' on the formation of his preferences (ibid.: 658).

While this reflects the general intuition that autonomous action is closely related to action that expresses someone's 'character', Buss now faces the challenge to capture the intuition that acting 'out of character' typically does not prevent someone from being held accountable for his actions (the terms 'character' and 'identity' will be used interchangeably in the following).[41] What makes an agent act autonomously, on Buss' view, is the fact that his preferences either reflect his character or otherwise were 'directly caused by a psychological and/or physiological condition that is not at odds with minimal human flourishing' (ibid.: 659). Buss relies heavily on what she refers to as the 'human flourishing condition' – implying that the (ultimately metaphysical) distinction between autonomy and heteronomy has to be informed by the normative notion of what makes a human life go well.[42]

---

[38] See also Buss (1994) for a thorough critique of the popular endorsement account of autonomous action.

[39] Buss talks about 'intentions', but let's stipulate that nothing of substance is lost when we translate that into the term 'preferences', which is economese for 'motives'.

[40] For Buss, autonomous action is necessarily 'self-determination in the passive mode' (ibid.: 657). She sees passivity 'at the heart of all agency', for 'nothing active can sustain its activity without passively relying on everything that makes this activity possible' (ibid.: 658, see also 684) – a striking echo of Sunstein's point that it's the choice architecture surrounding us that makes agency possible (see sect. 2, above).

[41] See (ibid.: 663). To be sure, the relevance of (more or less stable) character traits for explaining behavior is contested (see, e.g. Sreenivasan 2002), but that particular debate is of little relevance to our argument.

[42] Buss (2012: 672-685) elaborates upon this condition: She defines a 'trait' as interfering with autonomous agency if it is the case that 'when it is a stable disposition, it typically prevents [human beings] from satisfying one or more basic needs without exceptional effort' (ibid.: 672), where terms like 'basic needs' and 'exceptional' are deliberately left imprecise in order to do justice to the 'variability among healthy human beings' (ibid.).

Accordingly, Buss relaxes the condition (popular with standard accounts, see above) that to be autonomous – i.e., to play a decisive role in one's own preference formation –, requires a special way to exercise one's agential capacities. In contrast, she appeals to the 'identity all rational agents have insofar as they *are* representatives of their species, with their own characteristic way of functioning' (ibid.: 659, FN 29, emphasis added).[43]

Hence, an agent fails to act autonomously, on this view, only if his preferences reflect nonrational influencing factors *of a particular kind*, to wit, those that are 'elements or symptoms of human malfunctioning': Consider someone in the grip of mania, some self-destructive compulsion, or 'overpowering' emotions of rage: It seems intuitive to attribute preferences formed under conditions like these to something 'external' to the agent herself ('he's not himself') and, consequently, to exempt that agent from being held fully accountable for his ensuing actions (ibid.: 661).[44] Debilitating conditions can be seen as impediments to autonomy, for there's a sense in which sickness, depression and the like are 'hostile takeovers' of the self (ibid.: 668): they are deformations ('malfunctionings'), 'incompatible with a human being's identity as a representative of her kind', that therefore call for special care and treatment (ibid.: 669). When we are afflicted with conditions of this kind, we are not directly involved in the production of our own actions (ibid.: 670).

It's worth noting that Buss' account seems to be the only one among those discussed so far that takes seriously the intuition that self-government ('autonomy') is ultimately about a special way of relating *to oneself* – as opposed to relating, say, to Reason (ibid.: 666). Buss' account is quite undemanding in proposing that 'when our intentions are determined by physiological and psychological states that do not typically prevent human beings from functioning minimally well', we act autonomously. On Buss' view, the – possibly highly manipulative – manner in which an agent has *acquired* his character or personality does not compromise his autonomy, as long as the agents' preferences subsequently reflect the 'decisive influence of their character' (ibid.: 688). Analogous reasoning applies to the notorious case of the indoctrinated woman displaying adaptive preferences:[45] Unless she is prevented from functioning minimally well as a human being, her autonomy is not compromised (ibid.: 689; see also Buss 1994: 96).

---

[43] This shows, again, that Buss' account is wholly compatible with our EC criterion (see above).

[44] In other words: 'Under such circumstances, pathology takes the place of character' in explaining the agent's behavior (Buss 2012: 673). Buss points toward the asymmetry apparent when we are held accountable for actions reflecting, say, extreme joy, but not necessarily for those reflecting extreme depression (ibid.: 661-663); see Pizarro et al. (2003) for supporting evidence.

[45] This is reminiscent of Sen's famous case of the dominated housewife (Sen 1987: 45f.).

What, then, follows from this? Quite obviously, *genuine public nudging will hardly ever have a compromising impact on autonomy in Buss' sense*. On the contrary, to the extent that it succeeds in facilitating certain choices, it may help prevent agents from falling into debilitating conditions, thereby enhancing their autonomy. We submit, then, that contrary to what most critics claim, autonomy is hardly what's at stake in the ethics of nudging. As a corollary, we should not worry about the impact of 'manipulative' nudging on agents' self-government per se.[46]

Still, that's not the whole story, though. Buss (2012) refers to the basic compatibilist point that 'nothing can do anything unless it already is something, and that to be something is to have characteristic ways of behaving' (ibid.: 686). You need some identity in order to act at all. As we will argue in the following, nudging may be problematic in compromising the way agents form their identity in the first place. Put differently: *Given a personal identity*, genuine nudging would do little harm – but outside that assumption, it may be problematic. Here's why.

4.2 Action as self-constitution, not choice production

Analyzing the ethics of nudging under the assumption that agents have given identities is ultimately pointless. Nudging quite obviously influences agents' processes of preference (and, hence, identity) formation, which makes it necessary, when studying the ethics of nudging, to 'endogenize' those processes. Buss' account of autonomy, sketched above, can be seen as a first step toward that goal: On her view, what's crucial about autonomy is that an agent is able to let his character effectively shape his preferences – as we have seen, that ability can be negatively affected by debilitating circumstances such as mania. Buss however neglects the question of how agents acquire their character in the first place – conceding only that the way they do so may, under extreme circumstances, be inimical to autonomy qua violating the minimal human flourishing condition (Buss 2012: 688).[47] How, then, can we endogenize character formation? To be sure, some contributions to the ethics of nudging target its

[46] Note that this holds for public nudging (which we assume to be transparent, see above). Private nudging may be 'manipulative' by virtue of operating 'behind people's backs'. We would however only consider an agent to act heteronomously – given her information – if she were in the grip of Buss' debilitating conditions.

[47] As she puts it, in the unlikely event (read: thought experiment involving hypnosis or brainwashing) that some 'mad scientist blocks whatever relatively stable psychological dispositions are constituents of [an agent's] identity', this undermines autonomy, for 'having no ... psychic identity of any kind ... is incompatible with being a minimally well-functioning adult human being' (ibid.: 687f.).

character-moulding effects as well, albeit somewhat sparsely.[48] In order to incorporate this aspect more systematically into the ethics of nudging, let's see how James Buchanan and particularly Christine Korsgaard have elaborated upon a theory of agency that focuses on the problem of character formation.

Let's start with Buchanan. Famously, he rejects a justification in terms of the individual's epistemic privilege of the liberal order in general and of autonomy ('consumer sovereignty' in economese) in particular by arguing that '[t]he "individual", as described by a snapshot at any given moment, is an artifactual product of choices that have been made in prior periods, both by himself or herself and others' (Buchanan 1999b: 287). He elaborates upon this notion in his *Natural and Artifactual Man*, where he hypothesizes that 'what is special about human beings is our sense of "becoming", i.e. of becoming different from what we are' (Buchanan 1999a: 247). In our roles as artifactual beings, we partly construct ourselves, i.e. our own character. In economic terms, then, choice can be understood as an investment in becoming someone different (ibid.: 252). This goes beyond the well-known fact that agents may use preferences in order to *signal* some desired identity to others (consider your latest purchase of an Apple device). This artifactual nature of man grounds the notion that adult human beings are expected to assume responsibility for their own choices (ibid.: 257).[49] Buchanan's conception of individuality is strikingly reminiscent of what Davis (2014) refers to as 'bounded individuality' in the 'old' behavioral economics literature (notably Simon).[50]

Christine Korsgaard (2009) has elaborated upon a closely related account of agency. She starts from the observation that the human animal is unique in facing the existential challenge to *act* and, hence, to find reasons in order to guide her actions – as opposed to blindly following instinct, say. Human beings face the task of 'making something' of

---

[48] See, for instance, Bovens (2009: 214f.) on the risk that nudging 'infantilizes' people or 'fragments' their selves. Wight (2013: 130-32) comes closest to account for the importance of character formation in his critique of nudging. Sometimes, Mill's arguments on the value of active choice in promoting learning ('experiments in living') and building character is adduced in this context (e.g., ibid., Binder 2014). But see FN 51, below.

[49] Rothenberg (1962) has been one of the very few economists trying to elaborate upon the notion of consumers' sovereignty from such a viewpoint – when arguing, for instance, that '[c]onsumers' choices may not reflect their true tastes; but … maybe these tastes cannot accurately be known; or … they are not really "owned" but only "loaned" tastes anyway, passed on from one person to another. *What really can belong to the self and be accurately known is the experience of making and taking responsibility for choices*, whether right or wrong, and seeking to know by this continuing dialogue across the permeable boundary of the self what if anything is worth preserving. It is possible that this quest, given any reasonable degree of responsiveness in the outside world, is what consumers want more than being given what they are told they really want.' (ibid.: 282f., emphasis added)

[50] Davis (2014: 8f.) describes Simon's account of individuality as 'bounded' in the sense of being co-determined by external environmental factors. By interacting with their environment, agents reconstitute themselves over time. In contrast, in Kahneman's 'new behavioral economics', individuality is elusive: the individual risks elimination, due to inconsistent preferences, through either being money-pumped out of existence or fragmenting into multiple selves.

themselves (ibid.: xii), i.e., of constituting their identity through active choices. This life-long personal project is also the source of normativity – as Korsgaard argues, it presents humans with the formidable task to create and maintain *psychic unity* which makes actual agency possible. Put differently, action (as self-constitution) is tantamount to the project of achieving and maintaining a unified will or personal integrity over time. As Korsgaard puts it, '[w]e 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' (ibid.: 23f). An action of mine can be considered to be 'my own' (to be expressive of my identity) to the extent that it results 'from my entire nature *working* as an integrated whole' (ibid.: 18f., emphasis added). Note the procedural understanding of character in this quote.

It is important to stress the difference of this Buchanan-Korsgaard (henceforth BK) view to the approach, dominant in modern ethics (and also subscribed to by Buss, see above), which understands action as the production of something 'external' such as a choice, that is then subject to evaluation. As Korsgaard argues, that approach has been the one favored by Mill (ibid.: 8f.). An action would, then, be judged good to the extent that the choice – or, in Mill's case, also the preferences – produced by it were judged good.[51]

From the perspective advocated by BK, though, action is self-constitution, i.e., a person's 'identity' is not simply a state that has to be achieved first and from which action then issues, but a continuing process (ibid.: 43f.).[52] Then, however, it becomes obvious that the two Kantian imperatives of practical reason, viz., the Categorical and the Hypothetical Imperative, obtain compelling normative force for any human being struggling to maintain psychic unity. Hence, an action – comprising, as a package, both a specific act *and* a specific end (ibid.: 11f.) – would fail to promote the agent's psychic unity or integrity if it was either expressive of 'alien forces' or instrumentally non-effective (ibid.: 84). It is expressive of 'alien forces' to the extent that the agent fails to 'be something over and above the forces working within her' (ibid.: 134) – due to lack of principles. In that case, she lacks a unified will and her mind can best be characterized – paraphrasing Aristotle – as a mere 'heap of premises' (ibid.: 67). Korsgaard (2009: 168f.) illustrates this case with fictitious college student 'Jeremy', who, being unable to shape his will, stumbles through life, blindly following accidental impulses (a 'play of incentives' inside the mind, see ibid.: 76).

[51] Hence, while Mill may be interpreted as putting emphasis on the character-building function of active choice (see above), his account of agency seems ultimately incompatible with the notion of 'improving character'.

[52] Note that this shows how Korsgaard's notion of psychic unity differs from Bovens' notion of coherence (see above). Bovens focuses on the overall structure of preferences at a given moment in time – Korsgaard, though, understands the quest for psychic unity procedurally: It may involve temporarily crazy-looking structures, but make sense eventually as expressing the idiosyncratic struggle to balance conflicting personal roles and responsibilities. In a sense, it may express the ongoing aim to realize 'subjective coherence' (Schubert 2015).

Obviously, nudging may contribute to a situation of 'excessive convenience' resembling Jeremy's predicament, where agents are systematically discouraged from 'shaping their will' and exercising active choice. Therein lies the key problem. Note that excessive outsourcing of self-government – or, rather, excessive outsourcing of *agency* – may result in a vicious circle, making the agent eventually unlearn the skills (required in a behavioral world, see above) to extend his cognition, thereby making him unable to navigate life. One possible causal link from excessive outsourcing to loss of agency has recently been suggested by Waldron (2014, emphasis added): 'What becomes of the *self-respect* we invest in our own willed actions, flawed and misguided though they often are, when so many of our choices are manipulated to promote what someone else sees (perhaps rightly) as our best interest?'[53] More generally, on the BK view, the key to clarifying the normative costs of nudging lies in the answer to the question *What is the value of active choice*? In the presence of nudging the answer to that question depends on whether we assume a person's 'character' to be given (as, e.g., in Valdman 2010). In a behavioral world, though, where our character is endogenous, active choice is valuable to the extent that it succeeds in maintaining the agent's psychic unity; and it does so to the extent that it results from principles that the agent follows when organizing the 'forces' inside her mind, ideally in line with Kant's principles of practical reason.

What does all this imply for the ethics of nudging? We have to see whether and when nudging affects people's motivation and ability to pursue the project of self-constitution described by BK, i.e., to engage in active choice.

4.3 Some implications

Being equipped with an account that, we submit, grasps the true normative costs of nudging, let's briefly outline its basic implication for nudging, and then elaborate upon four key implications, ranging from the most applied to the most abstract.

The insight that it's agency, rather than autonomy that's at stake in nudging gives us a framework in which to discuss the ethics of nudging in a way that avoids intricate notions such as 'manipulation'. At the operational level, we should rather focus on *convenience*: The optimal level of nudging in a given life domain – say, old-age savings, food intake, smoking – results from a trade-off. Let's simplify: On the one hand, there can be 'too little' (public) nudging, i.e., agents with more or less limited mental resources are left to their own devices

[53] Hence, Korsgaard's account can capture Mill's intuition that the ability to choose actively is a 'muscle' that can be weakened unless exercised on a regular basis (on this, see also Sunstein 2015a: 43).

when facing the choice architecture present at any given moment. That choice architecture may be harmful in terms of its effects on agents' welfare or, indeed, agency. It may involve private choice architects treating 'the minds of other people as a resource' to be exploited (McCrawford 2015: 13) – not necessarily by 'manipulating' them (though some deliberate deception may be involved, depending on the way private nudging is regulated), but by discouraging them from engaging in active choice. On the other hand, there can be excessive (public) nudging, i.e., agents with more or less limited mental resources find themselves in a 'nanny state' offering them to delegate most key decisions of life to some external body of expert nudgers: Agents risk losing their capacity to engage in active choice, overcome challenges, and self-regulate, and ultimately their identity as idiosyncratic persons.

Put differently, choice architecture is Janus-faced: While it's needed to constitute agency, it can also compromise it. To illustrate, consider the simplification of government agency forms or the correction of objectively erroneous risk assessments (measures that help agents make instrumentally effective choices) on the one hand, and, on the other hand, the re-arrangement of food in cafeterias, relieving agents from the burden of self-control.[54] Behavioral policy – concerned, as it is, with the redesign of choice architecture – needs to find the right balance between insufficient and excessive nudging. This balance is obviously domain-specific. Let's briefly elaborate upon some implications of the BK account of agency.

(1) To make our criterion operational, we need a yardstick that separates those cases or life domains where it seems worthwhile to allow some active choice to be discouraged in the present (for the sake of enhanced possibilities to choose in the future), from those domains where this might not be the case. Here's a suggestion: With respect to the paternalistic use of nudging (its non-paternalistic use seems to be, grosso modo, unproblematic anyway),[55] we could use Rawls' list of primary goods: These are goods every human being has reason to desire, whatever specific overall conception of the good life she otherwise pursues. Note that this would significantly restrict the realm of nudging deemed legitimate, compared to what LP suggests. There are certain domains where it's plausible to presume that a large majority of citizens benefits from contextual support – from partially 'outsourcing their agency' – in order to minimize the risk of severe distress later in life. Thus, for instance, nudges might be considered to be less of a threat to agency overall when it comes to basic retirement savings or

[54] A somewhat trivial example is provided by GPS (a non-nudge, see above): It's obviously helpful in promoting agents' instrumental effectiveness (one of Korsgaard's conditions for successful agency), but may also erode agents' capacity to shape their own will when – literally – navigating the world (e.g. www.bostonglobe.com/ideas/2013/08/17/our-brains-pay-price-for-gps/d2Tnvo4hiWjuybid5UhQVO/story.html (accessed october 11, 2015).

[55] See Korobkin (2009) and Nagatsu (2015) on 'social nudges' that encourage the provision of public goods.

severe health risks.[56] On the other hand, the balance between normative costs (in terms of agency) today and potential benefits tomorrow may look differently in those domains where people's preferences are heterogeneous: Consider post-mortem organ donation, savings beyond some basic level, but also Thaler and Sunstein's notorious cafeteria case.

(2) As we have argued in the Introduction, the ethics of nudging can best be examined outside the narrow confines of Libertarian Paternalism. In light of agency considerations, LP looks distinctly unattractive as a normative policy program. First, it assumes given preferences and, by implication, given characters, thereby turning a blind eye towards considerations regarding autonomy and agency. Second, by accepting *homo oeconomicus* as a normative role model, it paves the way toward a wide variety of 'behavioral market failures' (Sunstein 2014: 34) that prima facie call for government intervention, which goes far beyond anything relating to primary goods (see our point 1) – in other words, implementing nudges in the framework of LP may compromise agency to a problematic degree.[57] Here's a conjecture: The BK notion of endogenous character is rather compatible with the conception of 'ecological rationality' (see above) in that the latter accepts a wide variety of environmental factors or cues as productive contributors to successful agency, and arguably supports the procedural approach to identity (see footnotes 49 and 50, above).

(3) 'Manipulation' in the sense conveyed by most critics of nudging – adhering, as they do, to the 'super-agent' view of autonomy – is hardly what makes genuine (i.e., transparent) nudging problematic. Consider Wight (2013: 136), deploring that nudges manipulate choice by smuggling an 'outside element' (i.e., one making the agent act out of character) into the process of preference formation, through which it then feeds back on the character, distorting further choices, and so on. This view assumes a given character; but as we have seen, a person's character should be seen as an *ongoing process* that continuously absorbs 'elements', most of which originate in private nudging anyway. Arguing in terms of 'manipulation' leads us into a conceptual mess: Hundreds of academic articles on the metaphysics of 'autonomy' have arguably produced more heat than light on what exactly should count as manipulation. The concept does little to inform the ethics of nudging; it should be skipped in favor of a focus on *convenience*: What matters is whether an agent faces too much of it, i.e., whether

[56] More intrusive regulatory tools might of course be considered, but here we are only concerned with nudging.
[57] Recently, Sunstein seems to relax LP's use of homo oeconomicus as a normative role model: '[L]et's understand the term [welfare] to refer to whatever choosers think would make their lives go well' (Sunstein 2013), which seems a bit vague. He also disputes, without however elaborating, that LP privileges what Kahneman famously refers to as 'System 2' (ibid.).

he's discouraged, to an excessive extent, from exercising his capacity to engage in active choice.[58]

Answering that question may not be straightforward, though. Consider changing the default with retirement savings schemes, often taken to be a paradigm case of nudging. Suppose that it's done in a transparent way. Obviously, this nudge discourages active choice (in its specific domain). This may affect the agent's ability to 'shape his will'. On the other hand, it may be argued to facilitate the agent's self-constitution by virtue of helping to prevent financial distress in old age. Moreover, nudges in this particular domain don't threaten agents' *moral* integrity – as opposed to, say, the much-cited case of default change in the domain of post-mortem organ donation (Johnson and Goldstein 2013).

(4) Finally, consider a key advantage of the Buchanan-Korsgaard account: it's *robust*, in two respects. First, it does not violate our EC criterion – it can obviously be coherently applied in the behavioral world in which nudging is supposed to operate. What's more, it also satisfies a second criterion that has already briefly been mentioned at the end of subsection 3.1, namely, the criterion of *Independence of Epistemic Privilege* (IEP): We submit that in order to be useful in the context of nudging, any conception of agency (or autonomy, for that matter) must be independent of specific empirical assumptions on the individual agents' epistemic privilege relative to some third party. Specifically, it should not presuppose that the individual agent exposed to nudging – the 'nudgee' – is systematically better informed about his own preferences than some external 'nudger'. Note that in many of life's domains, algorithms steering behavior can already today be safely assumed to 'know' and predict human agents' preferences (and even their 'character') better than the agents themselves. Thus, it may not be utopian to imagine an external body issuing nudges that are technologically perfect in the sense that they can channel people's behavior in a way that's in line with their own deepest commitments, i.e., that improve agents' welfare, *as judged by themselves*.[59] The notion of mere 'autonomy' would then not suffice to signal the normative costs intuitively associated with such kinds of nudging.

The Buchanan-Korsgaard account of agency, though, allows us to make sense of these costs: People subject to technologically perfect nudging would risk losing their motivation and ability to engage in active choice, i.e., self-constitution. They would risk losing their identity as recognizable human agents, rendering the very notion of technologically perfect nudging ultimately self-defeating. By dropping the assumption of a 'given character', we can

[58] The notion of active choice may also be easier to measure than whether someone's actions are unduly influenced by 'outside elements'.
[59] Consider the notion of 'smart' (personalized) nudges introduced in Smith et al. (2013).

remain unimpressed with the promises of technological progress in the realm of behavioral algorithms.

5. CONCLUDING REMARKS

The debate on the ethics of nudging (and public policy's cost-benefit calculus hopefully informed by it) should focus on their normative costs in terms of agency, rather than autonomy. That's no academic hairsplitting – the reorientation has important implications for the assessment of behavioral policies more general. Most importantly, when it comes to genuine (read: transparent) nudging we should be worried about the way it discourages active choice, rather than about its 'manipulative' nature. Put differently, the downside of nudging – and behavioral policies more general – may not be its 'distorting' or 'deceiving' impact on people's preferences, but rather the 'excessive convenience' it may create, a kind of harm neglected in Mill's classic 'harm principle' (Mill 1859: 21f.). By discouraging active choice, nudges may discourage people from engaging in the existential (if effortful) task of creative self-constitution that is at the heart of the very process of preference formation.

The notion that it's agency that's at stake in nudging may give us reason to explore questions that seem highly relevant for the ethics of psychologically informed behavioral interventions (an issue whose future relevance can hardly be overestimated), yet are still largely neglected in the literature. Should public policy – faced, as it is, with individuals displaying inconsistent preferences – abandon its focus on advancing people's 'true' preferences (or, equivalently, promoting 'true' utility) in favor of trying to maintain the complex institutional conditions necessary for agents' ongoing quest for self-constitution?[60] That perspective gives rise to other question: For instance, what's the contribution of *agency-friendly nudging* to social justice, considering that the poor seem to face particularly high cognitive burdens as they navigate the world (Mullainathan and Shafir 2013)? What's the relevance of moral character formation in the ethical assessment of behavioral interventions? Might the capacity of self-control, so famously cherished by Adam Smith (Bovens 2009), be particularly worth promoting? It seems that the project of incorporating psychological insights into economics has made it necessary to also reinvigorate the exchange between economists and moral philosophers.

[60] This step would relieve us from drawing (ultimately arbitrary) lines between 'legitimate' and 'illegitimate' wants or preferences. Preferences could be seen as just a contingent input into the idiosyncratic trial-and-error-based project of constituting and re-constituting one's identity. These thoughts require more reflection, though.

References

Akerlof, G.A. and Shiller, R.J. (2015) *Phishing for phools: The economics of manipulation and deception*. Princeton: Princeton University Press.

Bar-Gil, O. (2012) *Seduction by contract*. Oxford: Oxford University Press.

Barton, A. and Grüne-Yanoff, T. (2015) From libertarian paternalism to nudging – and beyond. *Review of Philosophy and psychology*, forthcoming.

Baumeister, R.F. and Tice, D.M. (1998) Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology* 74: 1252-1265.

Berg, N. (2014) The consistency and ecological rationality approaches to normative bounded rationality. *Journal of Economic Methodology* 21: 375-395.

Bhargava, S. and Loewenstein, G. (2015). Behavioral Economics and Public Policy 102: Beyond Nudging. *American Economic Review, Papers & Proceedings* 105: 396-401.

Binder, M. (2014). Should evolutionary economists embrace libertarian paternalism? *Journal of Evolutionary Economics* 24: 515-539.

Binder, M. and Lades, L.K. (2015). Autonomy-enhancing paternalism. *Kyklos* 68: 3-27.

Bovens, L. (2009). The ethics of nudge. In *Preference change: Approaches from philosophy, economics and psychology*, ed. T. Grüne-Yanoff and S.O. Hansson, 207-220. Berlin: Springer.

Buchanan, J.M. (1999a). The foundations for normative individualism. In his *The logical foundations of constitutional liberty, Vol. I*, ed. 281-291. Indianapolis: Liberty Fund.

Buchanan, J.M. (1999b). Natural and artifactual man. In his *The logical foundations of constitutional liberty, Vol. I*, ed. 246-259. Indianapolis: Liberty Fund.

Buss, S. (1994). Autonomy reconsidered. *Midwest Studies in Philosophy* 19: 95-121.

Buss, S. (2005). Valuing autonomy and respecting persons: manipulation, seduction, and the basis of moral constraints. *Ethics* 115: 195-235.

Buss, S. (2012). Autonomous action: Self-determination in the passive mode. *Ethics* 122: 647-691.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T. and Rabin, M. (2003). Regulation for conservatives: behavioral economics and the case for 'asymmetric paternalism'. *University of Pennsylvania Law Review* 151: 1211-1254.

Chetty, R. (2015). Behavioral Economics and Public Policy: A pragmatic perspective. *American Economic Review, Papers & Proceedings* 105: 1-33.

Christman, J. (1988). Constructing the inner citadel: recent work on the concept of autonomy. *Ethics* 99: 109-124.

Christman, J. (2005). Autonomy, self-knowledge, and liberal legitimacy. In *Autonomy and the challenges to liberalism*, ed. J. Christman and J. Anderson, 330-357. Cambridge: Cambridge University Press.

Conly, S. (2014). *Against autonomy: justifying coercive paternalism*. Cambridge: Cambridge University Press.

Davis, J.B. (2014). Bounded rationality and bounded individuality. Working Paper 2014-03, Dep. of Economics, Marquette University.

Deci, E.L. and Ryan, R.M. (2000). The 'what' and 'why' of goal pursuits: human needs and the self-determination of behavior. *Psychological Inquiry* 11: 227-268.

De Marneffe, P. (2006). Avoiding paternalism. *Philosophy and Public Affairs* 34: 68-94.

Dworkin, G. (2014). Paternalism. The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/paternalism/.

Elster, J. (1982). Sour grapes – utilitarianism and the genesis of wants. In Utilitarianism and beyond, ed. A.K. Sen and B. Williams, 219-238. Cambridge: Cambridge University Press.

Feinberg, J. (1986). Harm to self. New York: Oxford University Press.

Felsen, G. and Reiner, P.B. (2015). What can neuroscience contribute to the debate over nudging? *Review of Philosophy and Psychology*, forthcoming.

Felsen, G., Castelo, N. and Reiner, P.B. (2013). Decisional enhancement and autonomy: public attitudes toward overt and covert nudges. *Judgment and Decision Making* 8: 202-213.

Fischer, M. and Lotz, S. (2014). Is soft paternalism ethically legitimate? The relevance of psychological processes for the assessment of nudge-based policies. Working Paper, Cologne Graduate School, version may 9.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy* 68: 5-20.

Frey, B.S., Benz, M. and Stutzer, A. (2004). Introducing procedural utility: Not only what, but also how matters. *Journal of Institutional and Theoretical Economics* 160: 377-401.

Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology*, forthcoming.

Glaeser, E.L. (2006). Paternalism and Psychology. *University of Chicago Law Review* 73: 133-156.

Grüne-Yanoff, T. (2012). Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare* 38: 635-645.

Grüne-Yanoff, T. and Hertwig, R. (2015). Nudge versus boost: how coherent are policy and theory? *Minds and Machines*, online first, doi: 10.1007/s11023-015-9367-9.

Hagman, W., Andersson, D., Västfjall, D. and Tinghög, G. (2015). Public views on policies involving nudges. *Review of Philosophy and Psychology*, forthcoming.

Hansen, P.G. (2016). The definition of nudge and libertarian paternalism – does the hand fit the glove? *European Journal of Risk Regulation* 1/2016: 1-22.

Hansen, P.G. and Jespersen, A.M. (2013). Nudge and the manipulation of choice. *European Journal of Risk Regulation* 3: 3-28.

Harsanyi, J.C. (1982). Morality and the theory of rational behavior. In *Utilitarianism and Beyond*, ed. A. Sen and B. Williams, 39-62. Cambridge: Cambridge University Press.

Hausman, D.M. and Welch, B. (2010). Debate: to nudge or not to nudge? *Journal of Political Philosophy* 18: 123-136.

Hayek, F.A. (1961). The non-sequitur of the 'dependence effect'. *Southern Economic Journal* 27: 346-348.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Johnson, E.J. and Goldstein, D. (2003). Do defaults save lives? *Science* 302: 1338-1339.

Kirchgässner, G. 2015. Soft paternalism, merit goods, and normative individualism. *European Journal of Law and Economics*, forthcoming.

Korobkin, R. (2009). Libertarian Welfarism. *California Law Review* 97: 1651-1685.

Korsgaard, C.M. (2009). *Self-Constitution – Agency, Identity, and Integrity*. Oxford: Oxford University Press.

Lades, L.K. (2013). Impulsive consumption and reflexive thought: nudging ethical consumer behavior. *Journal of Economic Psychology* 41: 114-128

Le Grand, J. and New, B. (2015). *Government Paternalism*. Princeton: Princeton University Press.

Loewenstein, G. and Haisley, E. (2008). The economist as therapist: methodological ramifications of 'light' paternalism. In The foundations of positive and normative economics, ed. A. Caplin, A. Schotter, 210-148. Oxford: Oxford University Press.

Loewenstein, G., C. Bryce, D. Hagman, and Rajpal, S. (2014). You are about to be nudged. Working Paper, ssrn.com/abstract=2417383

McCrawford, M.B. (2015). *The world beyond your head – on becoming an individual in an age of distraction*. New York: Farrar, Straus and Giroux.

Mill, J.S. (1859). *On liberty*. Oxford: Oxford University Press.

Mills, C. (2013). Why nudges matter: A reply to Goodwin. *Politics* 33: 28-36.
Mills, C. (2015). The heteronomy of choice architecture. *Review of Philosophy and Psychology*, forthcoming.
Mongin, P. and Cozic, M. (2014). Rethinking nudge. HEC Paris Research Paper No. ECO/SCD-2014-1067, http://ssrn.com/abstract=2529910
Mullainathan, S. and Shafir, E. (2013). *Scarcity: Why having too little means so much*. New York: Time Books.
Nagatsu, M. (2015). Social nudges: their mechanisms and justification. *Review of Philosophy and Psychology*, forthcoming.
Nozick, R. (1974). *Anarchy, State, and Utopia*. New York: Basic Books.
Pal, G.C. (2007). Is there a universal self-serving attribution bias? *Psychological Studies* 52: 85-89.
Pichert, D. and Katsikopoulos, K.V. (2008). Green defaults: information presentation and pro-environmental behavior. *Journal of Environmental Psychology* 28: 63-73.
Pizarro, D., Uhlmann, E. and Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: the role of perceived metadesires. *Psychological Science* 14: 267-272.
Rebonato, R. (2012). *Taking Liberties*. London: Palgrave Macmillan.
Rothenberg, J. (1962). Consumers' sovereignty revisited and the hospitability of freedom of choice. *American Economic Review, Papers and Proceedings* 52: 269-283.
Schnellenbach, J. (2012). Nudges and norms: The political economy of libertarian paternalism. *European Journal of Political Economy* 28: 266-277.
Schnellenbach, J. (2016). A constitutional economics perspective on soft paternalism. *Kyklos* 69: 135-156.
Schnellenbach, J., and Schubert, C. (2015). Behavioral Political Economy: A Survey. *European Journal of Political Economy* 40 (B): 395-417.
Schubert, C. (2014). Evolutionary economics and the case for a constitutional libertarian paternalism. *Journal of Evolutionary Economics* 24: 1107-1113.
Schubert, C. (2015). Opportunity and preference learning. *Economics and Philosophy*, 31: 275-295.
Schubert, C. (2016). Green nudges: Do they work? Are they ethical? Working Paper.
Selinger, E. and Whyte, K. (2011). Is there a right way to nudge? The practice and ethics of choice architecture. *Sociology Compass* 5: 923-935.
Sen, A.K. (1987). *On ethics and economics*. Oxford: Basil Blackwell.
Smith, V. (2003). Constructivist and ecological rationality in economics. *American Economic Review* 93: 465-508.
Smith, N.C., Goldstein, D.G., and Johnson, E. (2013). Choice without awareness: ethical and political implications of defaults. *Journal of Public Policy and Marketing* 32: 159-172.
Sreenivasan, G. (2002). Error about errors: virtue theory and trait attribution. *Mind* 111: 47-68.
Sugden, R. (2004). The opportunity criterion: consumer sovereignty without the assumption of coherent preferences. *American Economic Review* 94: 1014-1033.
Sugden, R. (2008). Why incoherent preferences do not justify paternalism. *Constitutional Political Economy* 19: 226-248.
Sugden, R. (2010). Opportunity as mutual advantage. *Economics and Philosophy* 26: 47-68.
Sunstein, C.R. (2012). The Storrs Lectures: Behavioral Economics and Paternalism. Working Paper, version 11/29/12, ssrn.com/abstract=2182619.
Sunstein, C.R. (2014). *Why nudge?* New Haven: Yale University Press.
Sunstein, C.R. (2015a). Nudging and choice architecture: ethical considerations. Discussion Paper No. 809, Harvard Law School.
Sunstein, C.R. (2015b). Fifty shades of manipulation. Working Paper (Version 18.2.15), ssrn.com/abstract=2565892.

Sunstein, C.R. (2015c). Nudges, agency, and abstraction: A reply to critics. *Review of Philosophy and Psychology*, forthcoming.
Sunstein, C.R. and Thaler, R.H. (2003). Libertarian paternalism is not an oxymoron. *University of Chicago Law Review* 70: 1159-1202.
Sunstein, C.R. and Thaler, R.H. (2006). Preferences, paternalism, and liberty. *Royal Institute of Philosophy Supplement* 59, 233-264.
Tannenbaum, D., Fox, C.R., and Rogers, T. (2014). On the misplaced politics of behavioral policy interventions. Working Paper. http://scholar.harvard.edu/files/todd_rogers/files/on_the_misplaced.pdf
Thaler, R.H., Benartzi, S. (2013). Behavioral economics and the retirement savings crisis. *Science* 339 (6124): 1152-1153.
Sunstein, C.R. and Thaler, R.H. (2003). Libertarian paternalism. *American Economic Review, Papers and Proceedings* 93: 175-179.
Thaler, R.H. and Sunstein, C.R. (2008). *Nudge: Improving decisions about health, wealth and happiness*. New Haven: Yale University Press.
Valdman, M. (2010). Outsourcing self-government. *Ethics* 120: 761-790.
Waldron, J. (2014). It's all for your own good. *New York Review of Books*, october 9.
Watson, G. (1987). Free action and free will. *Mind* 96: 145-172.
White, M.D. (2013). *The manipulation of choice: ethics and libertarian paternalism*. New York: Palgrave.
Wilkinson, T.M. (2013). Nudging and manipulation. *Political Studies* 61, 341-355.