UNIVERSITY OF STIRLING

Stirling Management School

Autonomy-enhancing paternalism

Martin Binder
Leonhard K Lades

Stirling Economics Discussion Paper 2014-09

September 2014

Online at

http://www.stir.ac.uk/management/research/economics/working-papers/

Autonomy-enhancing paternalism

Martin Binder[a,b], Leonhard K. Lades[c]

[a]*Bard College Berlin, Platanenstr. 24, 13156 Berlin, Germany*

[b]*Max Planck Institute of Economics, Evolutionary Economics Group, Kahlaische Str.10, 07745 Jena, Germany*

[c]*Behavioural Science Centre, Stirling Management School, University of Stirling, Stirling, FK9 4LA, UK*

(m.binder@berlin.bard.edu, l.k.lades@stir.ac.uk).

Abstract

We present a form of soft paternalism called "autonomy-enhancing paternalism" that seeks to increase individual well-being by facilitating the individual ability to make critically reflected, autonomous decisions. The focus of autonomy-enhancing paternalism is on helping individuals to become better decision-makers, rather than on helping them by making better decisions for them. Autonomy-enhancing paternalism acknowledges that behavioral interventions can change the strength of decision-making anomalies over time, and favors those interventions that improve, rather than reduce, individuals' ability to make good and unbiased decisions. By this it prevents manipulation of the individual by the soft paternalist, accounts for the heterogeneity of individuals, and counteracts slippery slope arguments by decreasing the probability of future paternalistic interventions. Moreover, autonomy-enhancing paternalism can be defended based on both liberal values and welfare considerations.

Keywords: libertarian paternalism, behavioral economics, autonomy, preference learning, welfare economics

---

I. INTRODUCTION

Behavioral economics has shown that individuals' decisions are influenced by the decision-making context and that individuals sometimes make decisions that are not in their best interest (Kahneman, 2011; Thaler and Sunstein, 2008). In particular when decisions are made by the automatic, intuitive "System 1" (Kahneman, 2011), individual decision-making is not always characterized by full rationality, perfect information processing, and complete self-control. Individuals routinely take decision-making shortcuts and decide based on heuristics (Tversky and Kahneman, 1974, Camerer, 2004). While many of these heuristics tend to work well in a wide range of contexts (Gigerenzer et al., 1999), the literature in behavioral economics has mostly focused on situations where biases and other distortions lead to non-optimal results (Conlisk, 1996; Rabin, 1998; Kahneman, 2003).

The insights of behavioral economics have led to the implementation of behaviorally-informed policy interventions that help individuals to make better decisions without reducing their freedom of choice. Since freedom of choice is maintained, these forms of paternalism are classified as "libertarian" (Thaler and Sunstein, 2003; Sunstein and Thaler, 2003; Thaler and Sunstein, 2008; Sunstein, 2014).[1] Libertarian paternalists argue that deviations from standard economic rationality prompt for the benign intervention by a social planner (in the words of Thaler and Sunstein (2008): a "choice architect"). This choice architect acts *paternalistically* by designing choices in a way that takes into account and even harnesses biases and heuristics to "nudge" individual decisions in directions that these individuals would consider to be welfare-promoting when cognitively reflecting about the decisions with sufficient information at hand. To be considered *libertarian*, nudges must not be coercive by limiting individuals' freedom of choice or, as Hausman and Welch (2010, p. 126) add, by making alternatives significantly more costly.

In this paper, we present an alternative form of soft paternalism called "autonomy-enhancing paternalism" (AEP). As its name indicates, AEP seeks to support individuals' ability to make autonomous decisions. We suggest that most other forms of soft paternalism do not put enough emphasis on the importance of autonomy defined as the ability to make critically reflected decisions (as also

[1] A different example of soft paternalism is "asymmetric paternalism" (Camerer et al., 2003).

demanded in Binder, 2014).[2] AEP acknowledges that behavioral interventions can – and typically will – change the strength of decision-making anomalies over time, and favors those interventions that improve, rather than reduce, individuals' ability to make critically reflected, unbiased, autonomous decisions. While other forms of soft paternalism aim to improve the outcomes of individual decisions by modifying choice contexts (Thaler and Sunstein, 2008), AEP suggests to use behavioral insights to improve the processes that underlie individual decision-making, thus potentially benefiting the individual well beyond single choice contexts. AEP has many advantageous characteristics: it takes individuals' autonomy seriously,[3] it prevents manipulation through behavioral policy interventions, it accounts for the heterogeneity of individuals, and it counteracts slippery slope arguments. Finally, AEP can be defended based on both liberal values and welfare considerations.

The paper is structured as follows. Section 2 proposes "autonomy-enhancing paternalism" as a new form of soft paternalism. We present examples of behaviorally informed policy interventions that are autonomy-enhancing, as well as interventions that are not. Section 3 discusses strengths and weaknesses of AEP. To illuminate the verbal discussion, section 4 illustrates some of our main points in a simple formal model. Section 5 concludes.

II. AUTONOMY-ENHANCING PATERNALISM

2.1. Definition

We define autonomy as "the capacity of a person critically to reflect upon, and then attempt to accept or change, his or her preferences, desires, values, and ideals" (Dworkin, 1988, p. 48). Using the language of behavioral economic dual-process theories (e.g., Kahneman, 2011), we understand autonomous decision-making procedurally, viz. as being closely related to decision-making in the reflective System 2. An autonomous decision is made when the decision-maker has the possibility to let their reflective System 2 be responsible for the decision. This does not exclude the possibility that autonomous decisions are made by the intuitive System 1: individuals can deliberatively de-

[2] This paper is the result of extensive discussions between the authors that were triggered by the critical view on libertarian paternalism as expressed in Binder (2014), on which some of our criticism here build.
[3] See also Hausman and Welch (2010), Korobkin (2011), Mills (2013).

cide to let their System 1 influence the decisions (however biased these might be). But during an autonomous decision-process, individuals are always able to put System 2 back into power again. Also creating one's own decision environment, sometimes this is referred to as "self-nudging" (Lades, 2014, p. 115), is an act of autonomous decision-making in our definition. Autonomy is then the possibility to make critically reflected decisions in System 2 that are not hindered by external or internal forces. When an individual, for example, cannot resist the temptation triggered by the perception of a chocolate cake, the decision to feast on it is not an autonomous one in this definition.[4]

Autonomy-enhancing paternalism (AEP) suggests that behavioral interventions should foster critical thinking by strengthening System 2, weakening System 1, or encouraging that decisions are made in System 2 without affecting the strengths of the systems. Autonomy-enhancing policy interventions promote self-empowerment (see Mills, 2013) and aim to free individuals from irrelevant influences (see Hausman and Welch, 2010). AEP proposes that the benefit of behaviorally informed paternalism lies in its intention to help individuals overcome their biases and decision-making fallibilities and help them in making better thought-through, autonomous choices. AEP is somewhat akin to de-biasing strategies and related to Larrick's (2004) argument that equipping individuals with mental strategies is preferable to changing choice contexts, because these strategies can increase individuals' ability to apply newly learned skills in other decision contexts. However, AEP is different from using de-biasing strategies, as it suggests using behavioral insights to modify the choice architecture in a way that promotes critical reflection. Similar to de-biasing, the focus of AEP is on helping individuals to become better decision-makers; it aims to improve well-being through improving the *processes* of decision-making. This is in contrast to other forms of soft paternalism that aim to improve the *outcomes* of decision-making processes without concerning themselves with how the decisions come about. While interventions that help individuals to make good decisions are likely to be beneficial for the individuals, interventions that change the choice architecture to help individuals to become good decision-makers, who are able to use their

[4] A referee points out that a different form of autonomy would consist in the tempted individual doing sports after eating the cake. In this perspective, one can debate whether the eating of the cake is actually an autonomous decision. In our framework, it would not be (if done in the ways described), whereas the second decision to run off the dietary effect of the cake would be an autonomous decision. In an overall assessment of the whole decision-making sequence, we would argue that the episode has been partly autonomous. These sorts of considerations will become relevant later on, when we put our concept into an explicitly dynamic perspective.

System 2 to make critically reflected decisions, are likely even more beneficial for the individuals.[5]

AEP takes an explicitly dynamic perspective and acknowledges that behavioral interventions can have effects that last over time. From this dynamic perspective, improving decision-making processes, rather than just improving the outcomes of decisions, is particularly valuable. When behavioral interventions strengthen individuals' decision-making abilities over time, their welfare is likely to benefit beyond the time and domain where the interventions are effective. Behavioral interventions can influence individuals' abilities to learn about both their cognitive biases and their preferences. When individuals critically reflect upon their decisions, chances are that cognitive learning occurs. AEP prefers cognitive learning over non-cognitive learning because the latter often happens without the individual being aware of it and is thus more open to manipulation and the influence of other parties. AEP encourages those behavioral interventions that help individuals to become better decision-makers and thus make better informed, less biased, and more autonomous choices over time that may better reflect their true preferences.

To fix ideas, let us compare autonomy-enhancing policy interventions with other interventions by means of a simple illustrative formalization. Assume that individuals are endowed with an income of I and have to allocate their income between a healthy good x_H and an unhealthy good, say junk food, x_J with prices p_H and p_J, respectively. The individuals have bounded rationality and sometimes make welfare-reducing errors. Let us describe a deviation from economic rationality by a bias η. This bias might occur when predictions and/or decisions are made intuitively in System 1.[6] The bias affects the individuals' decision-making processes so that the individuals' decision problem can be described by

$$\max\text{-}\eta\ U(x_H,x_J) \qquad \text{s.t.} \qquad p_H x_H + p_J x_J \leq I, \qquad (1)$$

where *max-η* describes an admittedly simplified "behavioral economic" description of individual decision-making. Adding the decision-making bias to the equation allows us to formalize the effects that behavioral interventions have on the processes of individual decision-making; unbiased

[5] Note that we use the same language of dual process theories that libertarian paternalists use. However, libertarian paternalists use System 2 preferences to define what makes individuals better off and thus affects their welfare. AEP suggests encouraging System 2 decision-making. Libertarian paternalism links System 2 to outcomes, we link it to processes.

[6] We use this error term as a catch-all for all types of deviations from standard economic rationality, which can represent failures to perceive alternatives as well as a failure to correctly evaluate them (see Mullainathan et al., 2012 for a related reduced-form approach).

decision-makers are not affected by any behavioral intervention. Consider a benevolent planner who aims to reduce the consumption of x_J. Assume that the planner can choose between four different types of policy interventions: a ban, a tax, a libertarian paternalistic nudge, and an autonomy-enhancing intervention. Banning x_J will change equation 1 to

$$\max\text{-}\eta\ U(x_H) \quad \text{s.t.} \quad p_H x_H \leq I. \qquad (2)$$

Taxing x_J will increase the product's price so that equation 1 becomes

$$\max\text{-}\eta\ U(x_H,x_J) \quad \text{s.t.} \quad p_H x_H + p_J^{tax} x_J \leq I, \qquad (3)$$

where $p_J^{tax} > p_J$. The two policy interventions obviously either reduce freedom of choice (in the case of the ban) or significantly change economic incentives (in the case of the tax).[7] Hence, they are not libertarian according to the libertarian paternalistic definition. A libertarian paternalistic nudge might introduce a new bias, say θ, in order to neutralize the other bias so that equation (1) becomes

$$\max\text{-}\eta{+}\theta\ U(x_H,x_J) \quad \text{s.t.} \quad p_H x_H + p_J x_J \leq I. \qquad (4)$$

As a result of the new bias θ (with $\theta = \eta$), the individuals' decisions are *as if* they were not prone to any bias, and the individuals consume the optimal amounts of x_H and x_J.[8] Note that if we had not added the possibility of decision-making biases (in this case η and θ) in the equation, the nudge would not have changed the formalization. Finally, an autonomy-enhancing behavioral intervention aims to improve individual decision-making over time by increasing awareness of η and potentially reducing the bias so that individuals become better decision-makers whose choices are not affected by any bias anymore. AEP is a form of soft paternalism that aims to use behavioral insights to support individuals to become decision-makers who can best be approximated by

$$\max U(x_H,x_J) \quad \text{s.t.} \quad p_H x_H + p_J x_J \leq I. \qquad (5)$$

Note that this formalization suggests that autonomous decision-making in System 2 is similar to the traditional economic assumption of rationally optimizing individuals.[9] We are aware that while errors and biases tend to occur more frequently in System 1 than in System 2, most proponents of

[7] The distinction between taxes and bans is not perfect: One could argue that bans do not reduce the choice set, but just increase the price of (then illegal) products. Also taxes do not only change incentives, but might also reduce individuals' budget so that their choice set is reduced.

[8] The universe of nudges is large. Introducing a mechanism that uses one bias to neutralize another is just one of many ways how nudges can affect individual behavior.

[9] A lot can be criticized about using "rational choice" as the normative benchmark of how individuals should behave (Berg et al., 2011, Binder, 2014). While we don't think that individuals will ever attain such strict rational choice abilities, in the present paper we argue that obtaining a comparatively more rational decision-making ability is conducive to the individuals' best interests.

dual process theories do not argue that System 2 choices are rational according to rational choice postulates (see Evans and Stanovich, 2013; Kahneman, 2011). At most System 2 choices can be considered to be closer to a rational benchmark of optimal behavior. Our argument thus solely posits that System 2 choices tend to be more autonomous than intuitive System 1 choices, and that equation (5) is the best approximation for autonomous behavior we have.

2.2. Examples

In general, interventions that reduce the effects of biases on individual decision-making are autonomy-enhancing: this can encompass "simplification", "de-biasing", and "reframing" of consumer choices (see Larrick, 2004, Trout, 2005). Other, slightly more coercive instruments apply as well: cooling-off periods and "required active choosing" (mandated choice), for example, increase autonomy by allowing individuals to reconsider their choices and by increasing the awareness that decisions can be made, respectively. Some behavioral interventions do not enhance, or might even reduce, autonomy. Examples of these interventions are defaults, neutralizing one bias by another one, and influencing preferences by anchors. We elaborate on these examples in the following.

One of the most obvious examples of an autonomy-enhancing intervention, and the least restrictive on liberty, is simplification. Choice architecture that is too complex can overwhelm individuals and thus activate heuristics and biases in System 1. Instead of reflectively answering correct questions in System 2, overwhelmed individuals might use System 1 to answer related questions that are easier to answer (Kahneman, 2011). Presenting information in simple and intuitive ways that help individuals to make well-informed System 2 decisions does increase autonomy. For example, in the U.S. the food pyramid was replaced by "myplate" as a much simpler way to illustrate the five food groups that are the building blocks for a healthy diet (see Sunstein, 2013).[10] With this information at hand, individuals are able to better make critically reflected decisions. This is in line with Loewenstein and Haisley (2008), who argue for some form of behavioral intervention especially in situations that are complex and where people lack experience in decision-making. It is also in line with one of the principles of good regulation, "make it simple", as Sunstein (2013) has introduced

[10] A different example for de-biasing by simplifying a decision problem is to present conditional probabilities in the form of natural frequencies (see Gigerenzer, 2011).

it.[11]

Another instrument in the toolbox of AEP, which is related to simplification, is to reduce choice sets to magnitudes that do not overwhelm individuals so that they fail to come to autonomous decisions (Camerer et al., 2003; Trout, 2005). Providing all choice options may provide the maximal freedom of choice, but it is likely to reduce individuals' ability to make critically reflected decisions when individuals are overwhelmed by their choice sets and hence procrastinate on the choices (and choose not to choose). For example, requiring a person to choose from thousands of mutual funds for their pension schemes has been shown to decrease choice rates as opposed to choice situations with only few mutual funds (Iyengar et al., 2003). Reducing choice sets to manageable sizes might encourage (boundedly rational) individuals to actively make decisions (instead of, for example, relying on default rules). Enabling individuals to make these initial decisions on simplified, preprocessed choice sets may facilitate individuals' ability to make more complex decisions later on and is thus autonomy-enhancing. To ensure that individuals can critically reflect upon the larger choice sets later on, initial reductions of the choice sets should be complemented by the information about the existence of the larger choice sets. Such information reduces the risk that the preprocessing of the choice sets discards alternatives that might be of interest to some individuals, and the individuals' ability to look beyond preprocessed choice sets is maintained.[12]

Some autonomy-enhancing interventions activate System 2 so that although biases in System 1 might still be present, these do not drive decision-making. Forcing people to make choices is coercive as it reduces their choice sets by removing the freedom not to choose. However, such coercion can also increase individuals' autonomy: mechanisms of required active choosing can bring to attention the possibility to decide in situations where this possibility would have not been obvious and probably neglected by the individuals (Hausman and Welch, 2010, p. 134). For example, when individuals are asked whether to save for retirement or not before being allowed to start their job, they are made aware of the possibility and necessity to save for old age and thus are encouraged to

[11] AEP is explicitly not in line with other principles of regulation that Sunstein (2013) introduces, such as "make it automatic" and "don't strain System 2".

[12] Giving information about the existence of larger choice sets also allows individuals to respond individually to the pre-processing of the choice sets. Especially when individuals are heterogeneous, providing such information seems desirable.

reflect on their respective preferences.[13] Cooling-off periods increase autonomy, too, because decisions made in the spur of the moment can be reversed after critical (re-) consideration so that the individual has the possibility to make a critically-reflected decision afterwards.

Making individuals aware of decision-making anomalies identified in behavioral economics and cognate disciplines can enhance individuals' autonomy, too. Only when individuals know that decision-making biases may have an influence on their current and future behavior, can they critically reflect upon whether they are willing to let their behavior be influenced by these biases. For example, informing people about the factors that lead to impulsive choices can help the individuals to organize their lives to either generate or prevent impulsive decisions (Lades, 2014). This provision of information can be combined with offering commitment devices that critically reflecting individuals can install to constrain their future selves' behaviors, which would be considered to be an instrument of AEP (see Bryan et al., 2010).

While many behavioral interventions discussed in the literature on soft paternalism can enhance autonomy, there are also interventions that do not enhance, or even reduce, autonomy. For example, many default rules are likely not autonomy-enhancing. If individuals are prone to inertia, have preferences for a status quo, and are subject to the "yeah-whatever heuristic" (Thaler and Sunstein, 2008), they might learn to rely on default rules set by the government or other choice architects. As a result, behavioral biases might become stronger over time. For example, in an experimental setting, de Haan and Linde (2012) show that good defaults can induce individuals to (mistakenly) follow random defaults later on. In a related vein, Carlin et al. (2013) show that the information inherent in defaults can reduce individuals' information acquisition incentives so that social welfare decreases.

Also behavioral policy interventions that aim to neutralize one decision anomaly by another one may reduce, or, at least, do not enhance, autonomy. Consider the use of risk narratives as suggested by Jolls and Sunstein (2006). These risk narratives aim to counteract individuals' optimism bias (which causes too much risk taking) by harnessing individuals' availability bias. The government could demand producers of risky products to provide the consumers with stories of single instances

[13] Too many of such forced choice paradigms, however, might overstrain the individuals. This would necessitate quick System 1 decisions and thus cause reductions in autonomy.

where the use of the product in question had problematic outcomes. Since individuals overemphasize such information, the narrative can neutralize individuals' optimism bias.[14] Another example is the organization of food in cafeterias as described in Thaler and Sunstein (2008). Choice architects utilize the fact that food products presented on eye level in cafeterias are purchased more often than food products in less salient positions (Hanks, et al. 2013) to neutralize present-biased preferences for tasty but unhealthy food items. Putting healthy food in the most salient positions is likely to induce more healthy consumption already in the short run. However, learning opportunities might be missed, and individuals' present-biased preferences are not reduced by the arrangement of the food. Whenever one bias is successfully used to neutralize another bias, the outcome of the decision is *as if* the individuals behaved in an unbiased way. However, the decision-making process that led to the decision was not an autonomous one.

Finally, setting anchors that the individual does not know about can reduce autonomy because anchors, like many nudges, can induce preferences of which the individual is not aware. Anchors can be used in cases where individuals are conjectured not to have well-formed preferences and where anchors function as a signal of what the policy-maker thinks would be a reasonably welfare-improving choice (e.g. pension savings rates of "2%…5%" vs. "10%…20%" of one's wages). If anchors are used without providing an explanation for the height of the anchor chosen, individuals are likely to not learn from making their choice between the preprocessed alternatives. Individuals might think that their expressed preference is their preference although the preference was induced in the instant before through the anchor. In the long run, this will lead to individuals forming preferences that were ill-reflected and can be considered heteronomous. In order to be an acceptable tool in the toolbox of AEP, anchors would always need to be complemented by some sort of cue that makes individuals aware that in their heterogeneous life circumstances, a different value than the anchor chosen by the policy-maker might actually be more appropriate (e.g. "Warning: retirement savings of 10% of your monthly wages will avoid poverty on average, but this amount needs to be increased depending on your age and existing savings.").

III. DISCUSSION

[14] Thanks to an anonymous reviewer for pointing at this example.

In the following, we discuss important features of AEP and compare them briefly to the currently most popular form of soft paternalism, libertarian paternalism (Thaler and Sunstein, 2008). While a subset of behavioral interventions discussed in the literature on soft paternalism are in line with both forms of paternalism, the behavioral intervention's justifications can differ between both approaches. Discussing on what grounds policy interventions are justified or not is important because it is likely that in the future more behavioral biases will be identified in behavioral economics and cognate disciplines and more behavioral interventions will be used (which need to be justified). Also for interventions that do not exist yet it is important to have clear guidelines to evaluate their effects, possible dangers, and permissibility.

3.1. Autonomy and freedom of choice

In the libertarian paternalistic logic, acceptable behavioral interventions are supposed to preserve freedom of choice (and do not significantly change incentives). Preserving freedom of choice, however, is a very narrow view of defining liberty (see e.g., Qizilbash, 2012, Korobkin, 2011).[15] Liberty can also be understood (more substantively) in terms of the ability to make critically reflected, i.e. autonomous, decisions. This ability is an important feature of liberal and libertarian views, and interventions in the spirit of AEP increase this ability.

One advantage of focusing on autonomy, rather than on freedom of choice, is that maintaining and even fostering individual autonomy acts as a safeguard against exploitation and manipulation of the individual by the paternalist. Interventions that encourage individuals to make critically reflected decisions are transparent. To be sure, libertarian paternalism does not imply that the individual will always be manipulated. But the core theoretical concept of libertarian paternalism does allow that some nudges are manipulative (this holds especially for default rules) and keep individuals in the dark about the fact that there is some behavioral regulation. Only by adding an additional

[15] Libertarian paternalism is also internally inconsistent when accepting the notion of liberty as freedom of choice. If individuals are "tricked" into choices by the libertarian paternalist through the setting of default rules, they might not be aware of being nudged into certain behaviors. When an individual is not aware of the nudge, only nominal freedom of choice is left intact, for all intents and purposes, the real choice set (the one the individual acts on) is decreased and choices are de facto "blocked" (see also Rebonato, 2012, p. 132; this is not to say that this holds for all cases and all tools in the toolbox of the choice architect, but it will likely hold for a number of tools). One positive thing to say in favor of hard paternalism is that individuals are at least aware that they are being curtailed in their liberty or autonomy and can try and oppose this. The hidden character of some of the tools of the libertarian paternalist make this extremely difficult and by this disable another safeguard against this particular slippery slope (Rebonato, 2012, p. 132).

"transparency criterion" (see Bovens, 2009) libertarian paternalists prevent the possibility of manipulation. AEP does not need an additional criterion to prevent manipulation. Already the focus on encouraging autonomous decision-making renders manipulation less likely. Thus, AEP is not susceptible to objections arguing that soft paternalism is not acceptable because manipulation is morally undesirable, contravenes liberal ideas, and paints a cynical picture of the role that legislators and policy-makers should play vis-à-vis sovereign citizens (this has been pointed out similarly by Hausman and Welch, 2010, p. 134).

Distinguishing between autonomy and freedom of choice also allows making trade-offs between both liberal values. Reducing individuals' freedom of choice can increase their autonomy and vice versa. Acknowledging the importance of autonomy allows justifying some interventions that benefit the individual even when freedom of choice is reduced. For example, although reducing the number of mutual funds that individuals have to choose from for their pension schemes (Iyengar et al., 2003) is often related to libertarian paternalism, the reduction of the choice set is not "libertarian" as defined in the concept. Our point here is that libertarian paternalism with its focus on freedom of choice is somewhat blind to the possibility that reduced freedom of choice in favor of increased autonomy might be a desirable trade-off.

3.2. Autonomy and the dynamic perspective

Many behavioral interventions have a strong intuitive appeal when described as one-shot situations without any regard for inter-temporal dynamics. In one-shot situations, one might be tempted to approve of many interventions that have the ability to correct a given choice and thus increase individual well-being, even at the expense of autonomy. However, acknowledging that interventions often also influence future choices can change the evaluation of these interventions, and dynamic effects need to be explicitly considered. We highlight two interdependent ways how behavioral interventions can affect future choices: (a) via preference learning and (b) via changing the strength of the deviation from rationality.

Regarding the behavioral interventions' effects on preference learning, it is important to acknowledge that individual preferences cannot be reasonably well assumed to be given and stable (Witt, 1991; Binder, 2010). Some actions are not only driven by preferences, but also create or

change preferences (Ariely and Norton, 2008). Economic processes and policies, including soft paternalistic interventions, can shape preferences. A behavioral intervention at time t+0 may (and likely will) influence choices at time t+1 and afterwards. For example, a behavioral intervention that increases individuals' propensity to eat healthy may reduce their propensity to eat junk food in the future because individuals get used to healthy eating. While this is a case where individuals might like their vanishing preference for junk food, one has to be aware of the effects of behavioral interventions on preference learning when evaluating soft paternalistic policy interventions.

Preference learning and change is well-researched in psychology: preferences can be learned in a cognitive way, i.e. autonomously as a result of cognitive reflection. But most often individuals learn preferences via associative learning without being aware of this (Hergenhahn and Olson, 1997; Witt, 2001).[16] In this case, behavioral interventions are more effective (also nudges work best "in the dark", see Bovens, 2009), but also more problematic. If behavioral policies induce unconscious preference learning, individuals are not able to make conscious and autonomous decisions about which preferences they want to learn. Without the ability to make autonomous decisions, behavioral interventions can put individuals at danger of losing their ability to pursue their own happiness (Schubert, 2012).[17] Moreover, preferences resulting from associative learning tend to be highly stable and difficult to unlearn, in parts because of the low conscious involvement, but also because of repeated reinforcement over long time horizons.[18] Even if, at one point in time, individuals were to become aware of the intervention's influence on their preferences, the stickiness of associative preference learning can actually make it very difficult to make a critically-reflected decision in the future and reverse this preference: while this obviously holds for addictive preferences (a benign example of which might be coffee), many other preferences that were learned without cognitive involvement and have been reinforced many times come to mind (food preferences, social preferences, etc., see Zajonc and Markus, 1982). Interventions in line with AEP encourage individuals to make conscious decisions about which preferences they want to learn;

[16] Many of our childhood preferences are acquired via associative learning and tend to be not easily reversible (Zajonc and Markus, 1982).

[17] Nudges can also put individuals at danger of "learning" helplessness (Binder, 2014). This can create preference learning trajectories where individuals are locked-in into preferences one can doubt are in the individuals' actual interests (see also Schubert and Cordes, 2013). These preferences are likely to reflect the norms society adheres to at a given moment (or, even worse, they reflect ad hoc goals of policy-makers).

[18] Whether preferences can be unlearned depends inter alia on the reinforcement schedule (see more extensively Binder, 2010, sec. 6.4.3); even if preferences can ultimately be unlearned, the costs associated with this are much higher than libertarian paternalists claim they would be in the case of opt-outs.

AEP does not try to influence individuals' preference learning paths under the radar of individual consciousness.

Behavioral interventions can also change the strength of decision-making biases over time. Most libertarian paternalistic nudges operate via System 1 without encouraging critical reflection and are thus unlikely to reduce decision-making biases. To the contrary, nudges might even strengthen decision-making anomalies over time. Individuals might implicitly learn to trust the choice architecture and to rely on their System 1 when making decisions. For example, defaults, especially good defaults, might strengthen individuals' inertia and make them inactive so that their future behavior is even more strongly influenced by future defaults (see de Haan and Linde, 2012; Carlin et al., 2013). This could encourage policy-makers to make more use of behavioral interventions over time. AEP favors interventions that promote, rather than reduce, individuals' ability to make critically reflected decisions in their System 2. If behavioral paternalistic interventions increase autonomy and facilitate critical reflection, future behavioral interventions will become less powerful over time. Individuals will make more and more decisions in their System 2, and these decisions are not prone to the biases that soft paternalistic interventions try to harness to influence behavior. The dynamic perspective inherent in AEP thus makes a case for less soft paternalism over time and counters reservations against soft paternalism based on slippery-slope arguments (Rizzo and Whitman, 2009).[19]

Sunstein (2014) proposes that it is hard to see why libertarian paternalistic nudges should be seen objectionable if they work only or largely because of the operations in System 1. Since choice architecture is inevitable, he argues, nudges should not be ruled off limits merely because they work as a result of the operations of System 1, as long as they are made public and defended on their merits. However, when considering that nudges can change the strength of decision-making anomalies over time, and influence the degree to which people make decisions in System 1 or System 2, one can see some additional costs of nudging. It seems to us that behavioral policy interventions – in particular those implemented by the government – should not enlarge the scope of the choice architecture that is "inevitable". Just like freedom of choice is a matter of degree (there can be

[19] Some critics also argue that oft paternalism can increase the probability of hard paternalism, thus creating a slippery slope with regard to the number and types of interventions (e.g., Rizzo and Whitman, 2009). We emphasize dynamic changes occurring within the nudged individuals, not within policy-makers or societal norms.

more or less freedom of choice), decisions can be more or less strongly influenced by the choice architecture, and autonomy-enhancing interventions make sure that the power of the choice architecture does not increase over time.

3.3. Autonomy, welfare, and the dynamic perspective

While we argue that behavioral interventions should enhance autonomy, ultimately, the overriding value judgment to justify paternalism has to be that interventions increase individual welfare (otherwise the interventions would not be paternalistic). From the perspective of AEP, however, it is not justified to increase welfare at the expense of autonomy. But it is also not justified to increase autonomy at high costs of welfare. Hence, the effects of behavioral interventions on individual autonomy and welfare must be evaluated. One argument in favor of libertarian paternalistic behavioral interventions is that their net effect on welfare is positive (Sunstein, 2014). On the one hand, large welfare gains can be realized by making decisions for less rational individuals that make them better off. On the other hand, the costs of these interventions are supposed to be low as they consist of only small reductions in freedom of choice and some behavioral influence is inevitable (e.g., Trout, 2005; Thaler and Sunstein, 2008; Camerer et al., 2003; Sunstein, 2014). However, behavioral interventions can affect individuals' welfare both in the short run and the long run, and both immediate and delayed costs and benefits have to be taken into account when engaging in a cost benefit analysis of behavioral interventions. Autonomy, we argue, is essential when analyzing the delayed effects of behavioral interventions on individual welfare.

Acknowledging that behavioral interventions at time t+0 can influence individual welfare also in the future (at time t+1) prompts a more careful distinction of cases, and it is an empirical question which behavioral interventions fall into what category. "Ideal autonomy-enhancing interventions", which AEP encourages, increase autonomy (and possibly also welfare) at time t+0 and through their dynamic learning effects, autonomy and welfare are also increased in the future at time t+1. "Painful autonomy-enhancing interventions" increase autonomy in t+0, but also reduce welfare in the short run, for example because learning costs are high, it takes effort to engage in active decision-making, or different cognitive biases offset each other (see Besharov, 2004).[20] Through in-

[20] A different, practical question would be how to deal with cases where several biases interact. Our hunch in the spirit of self-limiting paternalism (anti-slippery-slope) would be to restrict interventions to clear-cut cases where such inter-

creased autonomy, however, AEP suggests that welfare in t+1 may be increased. The evaluation of the success of these paternalistic interventions then depends on whether long-run welfare gains (in t+1) outweigh the short-run costs in t+0. The welfare gains in t+1 of interventions that enhance individuals' autonomy are particularly large when possibilities for learning exist and when similar decisions will be repeated in the future. When there is no possibility for individuals to learn, or when the decisions affected by the behavioral intervention are once-in-a-lifetime decisions, the positive long-run effects of autonomy-enhancing interventions on welfare are likely to be smaller (but even then, making an important decision in one domain might provide a lesson for other domains).[21]

"Manipulative, but well-motivated interventions" increase welfare in t+0, but undermine autonomy so that individuals cannot learn from the situation, and welfare in t+1 might be decreased. In the short run, such manipulative interventions that utilize the knowledge about biases and decision-making anomalies might be more effective in generating welfare-enhancing outcomes than interventions that encourage critical reflection. For example, when defaults are set, the short run effort of making decisions is smaller that when individuals are required to make their own choices so that the defaults may lead to higher welfare than mandated choice arrangements.[22] In the long-run, however, this relation might reverse. While it might be the case that default rules impose less costs for a given level of welfare gain at time t+0, they might come at the expense of welfare at time t+1: in the long run, default rules might lead individuals to learn to be dependent and inactive. Individuals then, without the constant attention of the paternalist, might suffer welfare-losses since they never learned to autonomously decide in such situations. Relieving individuals of the need to make complex and important life decisions will likely have higher long-term costs than a one-shot perspective suggests.[23]

---

action effects are sufficiently well-known to have an estimate of welfare effects in both periods.

[21] There is also the possibility that autonomy breeds misery", i.e. the "unhappy egg-head effect", where increasing autonomy leads to lowered welfare in both periods. Such interventions do not fall in the remit of AEP because AEP is a form of paternalism and hence aims to increase individual well-being.

[22] It might be the case that behavioral interventions impose other short-run costs on individuals. For example, when individuals derive utility from choosing, default rules that take away the recognition that a choice is necessary might reduce utility (Sunstein, 2014).

[23] A related dynamic aspect that is neglected in the one-shot perspective of libertarian paternalism is the question of the correct temporal interval to judge paternalistic interventions. Many behavioral interventions seem plausible when only the immediate effects on welfare are considered. But often it is appropriate to consider longer time intervals when evaluating the success of interventions. Consider, for example, an individual who binges on cake at time t+0. Assume that a regulator aims to combat this behavior and "defaults away" the cake so that the individual does not binge. But is

A final consideration when it comes to the welfare-aspects of AEP is the heterogeneity of actors: One of the arguments often used to criticize libertarian paternalistic policy interventions is that they do not take into account that individuals are heterogeneous: one-size-fits-all interventions are likely to reduce welfare of individuals who deviate from the average. For example, individuals might be defaulted into a pension saving program that does not fit their individual circumstances. AEP does not have this problem as its toolkit respects individual heterogeneity: no decisions are made for individuals, but individuals are made aware of the need to make a decision. Required active choosing, potentially combined with simplification procedures, for example, helps individuals to make a decision, but the ultimate decision is made by the individuals based on their individual idiosyncractic preferences and life circumstances (albeit based on a less biased decision-making process than without AEP).

3.4. Limitations

AEP is not without faults: maybe the strongest disadvantage of AEP comes in terms of a cost benefit analysis based on its welfare implications: interventions that encourage critical reflection in System 2 slow down the process of decision-making. Decision-making in System 2 uses more cognitive resources than fast and automatic decision-making in System 1. This loss in efficiency may reduce individuals' welfare. As Sunstein (2013) observes, people are busy and do not have much time to make a lot of decisions in System 2. Hence, he suggests following some general principles of behavioral regulation, including: make it simple, intuitive, meaningful, automatic, and don't strain System 2. While the first three principles are also in line with AEP, the last two are clearly not. AEP proposes that in order to reduce the strain on System 2 it is much better to make decisions easier so that System 2 can make autonomous and quick decisions than to make decisions automatic, let System 1 and all its biases decide, and possibly reduce the individuals' ability to make critically reflected decisions over time. Nevertheless, forcing individuals to make too many important autonomous decisions might lead to decision fatigue (Baumeister et al., 1998; Hagger et al., 2010) and hence has to be acknowledged as a disadvantage.

---

there the need for a nudge if the policy-maker could similarly inform the individual of the bad habit and suggest doing sports to offset the effect? A critically-reflecting individual might choose to binge on the cake in t+0 and go for an extended workout in t+1. When considering this extended interval of time it is unclear why there should be any intervention that tries to make the individual not eat the cake in t+0.

Secondly, autonomy-enhancing interventions could generate unforeseen consequences. For example, when multiple biases offset each other so that the outcomes of individual behavior are (nearly) optimal, autonomy-enhancing interventions that counteract only one of the cognitive biases might reduce individual welfare (see Besharov, 2004).[24] This is particularly problematic when policy-makers do not have the perfect information about the system of biases that might influence individuals' behavior. In such cases, AEP encourages interventions that reduce all biases; but even when only a subset of the biases can be reduced, AEP is in favor of that, as long as a long-run welfare gain is likely to result. This would be a case of "painful autonomy-enhancing interventions" as described in section 3.3. Note, however, that an autonomy-enhancing intervention that is not likely to increase welfare at least in the long-run would not be considered AEP, because such an intervention would lack an important qualifier of "paternalism", namely its aim to positively influence welfare.

Finally, in some cases the possibilities to critically reflect upon one's decision and to reverse one's choice can reduce individual welfare, because the possibility to reverse a decision can lead to decision regret. Individuals only fully commit to a decision when the decision is no longer reversible (Schwartz, 2000, Gilbert and Ebert, 2002) so that it is sometimes better for one's welfare not to have the opportunity to reverse one's choice. This case, to which a referee alerted us, shows a trade-off between welfare and autonomy. While AEP would suggest in this case that the paternalist is ill-advised to try to improve welfare through taking away the possibility to revert the choice after critical reflection, libertarian paternalism might promote such an autonomy-reducing, but welfare-enhancing intervention.

IV. ILLUSTRATION: "OPTIMAL SIN NUDGES"

In this section we present a simple formal model in order to illustrate effects of autonomy-enhancing interventions on individual well-being over time. The model is a variant of O'Donoghue and Rabin's (2006) model that analyzes optimal taxes for "sin goods" assuming that individuals have present-biased preferences. We additionally assume that individual perceptions of costs and

[24] We like to thank an anonymous reviewer for highlighting this limitation.

benefits of sin goods can be distorted. Our analysis is thus related to Besharov (2004) who investigates the effects of corrective interventions when individuals are prone to a combination of several biases (he considers present-biased preferences, regret, and overconfidence). We investigate an initial situation where present-biased preferences but no distorted perceptions are present and lead to overconsumption of sin goods. We assume that choice architects aim to reduce the consumption of these sin goods. We do not investigate harder forms of paternalism such as bans or taxes. In our illustrative case, behavioral interventions can either reduce present-biased preferences or operate through distorted perceptions. Since the initial overconsumption of sin goods is the result of present-biased preferences alone, interventions that reduce the present-bias enhance autonomy and interventions that change individuals' perceptions decrease autonomy. We sketch possible effects of these behavioral interventions on individual well-being, acknowledging that behavioral interventions can influence individual behavior for more than the contemporaneous time period.

4.1. The model

Assume that individuals (for example children in a school cafeteria) can choose between two types of goods: "normal" goods and "sin goods". The latter are particularly unhealthy (think of junk food). Normal goods are a composite of many goods. Sin goods generate utility with a decreasing marginal rate. Following Kahneman et al. (1997), we call this utility "experienced utility". The consumption of sin goods has negative consequences in the future. Consumers are endowed with income I, which is large compared to the consumption of the sin goods, and they do not save or borrow. The experienced utility from consuming the sin good x_t and the normal good z_t at time t is given by $u_t = v(x_t; \rho) - c(x_{t-1}; \gamma) + z_t$. While the function $v(x_t; \rho)$ represents the immediate benefits from current consumption of the sin goods, $c(x_{t-1}; \gamma)$ depicts the negative health consequences from past consumption of sin goods. The parameters ρ and γ are exogenously given context variables determining the extent to which the consumption of x_t leads to benefits and costs, respectively.[25] We assume that $v_{x\rho} > 0$ and $c_{x\gamma} > 0$ so that a higher ρ reflects a higher marginal benefit from consumption, and a higher γ reflects a higher marginal health cost from consumption. We normalize the price of the normal (composite) goods to be unity (our numeraire).

[25] In O'Donoghue and Rabin's (2006) model, these variables capture population heterogeneity in tastes.

We assume that the benefits and costs of period-t consumption are additively separable from the benefits and costs of consumption in any other period. Hence, for each (x, z)-bundle, the individual's experienced utility can be written as

$$u(x,z) = v(x; \rho) - \delta c(x; \gamma) + z, \quad (5)$$

where δ depicts the conventional discount factor. In the following, we will assume δ to be unity. In every period, individuals could maximize their experienced utility given by equation (5) by allocating their income I to the two goods x and z in an optimal way. The optimal allocation of income (x**, z**) maximizes experienced utility subject to the resource constraint $x + z \leq I$. Hence, x** satisfies $v_x(x^{**};\rho) - c_x(x^{**};\gamma) - 1 = 0$ and $z^{**} = I - x^{**}$.

Individuals' actual behavior, however, sometimes deviates from the behavior that maximizes experienced utility. Following Kahneman et al. (1997), we describe these deviations by assuming that individuals actually maximize their "decision utility" (or "wanting"), although maximizing their experienced utility (or "liking") would be optimal for their well-being (see also Besharov, 2004).[26] In this model, present biased preferences and distorted perceptions of the costs and benefits of sin goods can lead to dissociations between decision utility and experienced utility. Individual perceptions are described by α. When $\alpha = 1$, there is no distortion. However, when α deviates from unity, the perceived effects of the consumption of x on present benefits and future costs can be lower ($\alpha < 1$) or higher ($\alpha > 1$) than the true effects.[27] When integrating the possibility of distorted perceptions and assuming that δ equals unity, the decision utility u^d at time t can be described by $u^d(x, z) = v(x; \alpha_\rho\rho) - c(x; \alpha_\gamma\gamma) + z$.

Present-biased preferences describe the tendency to impulsively pursue immediate gratification. Following e.g. Laibson (1997) and O'Donoghue and Rabin (2003), we assume that the individuals' inter-temporal decision utility at time t is given by $U(\hat{u}_t,\ldots\hat{u}_T) = \hat{u}_t+\beta\sum^{T}_{\tau=t+1} \delta^{\tau-t}\hat{u}_\tau$, where $\hat{u}$ is the individual's decision utility at any given point in time and δ is a conventional discount factor which we assume to be unity. Present-biased preferences are described by the parameter β. While β

[26] Actual happiness is synonymous to "experienced utility" as introduced by Kahneman et al. (1997) and Berridge and Aldrige's (2008) "liking". We use actual happiness as benchmark for individuals' interests/well-being. Individual preferences are synonymous to "decision utility" (Kahneman et al., 1997) and "wanting" (Berridge and Aldridge, 2008). Preferences, but not actual happiness, can be subject to faulty affective forecasts or other biases (see more extensively Lades, 2012; Witt and Binder, 2013).

[27] We acknowledge that in many situations it is difficult to define the meaning of $\alpha = 1$ because some choice architecture is often inevitable.

is below 1 in terms of decision utility, in terms of experienced utility, β is equal to 1 (see Lades, 2012 for a micro-foundation of β based on the dissociation of utility into components of wanting and liking). Since we assume that also the perceived benefits and costs from period-t consumption are additively separable from the perceived benefits and costs from consumption in any other period, at any time period the decision utility $\hat{u}(x, z)$ corresponding to the consumption of x and z with the potential for distorted perceptions ($\alpha \neq 1$) and present-biased preferences ($\beta < 1$) can be written as

$$\hat{u}(x, z) = v(x; \alpha_\rho \rho) - \beta c(x; \alpha_\gamma \gamma) + z. \quad (6)$$

In every period, the individual chooses a consumption bundle (x, z) that maximizes equation (6) subject to the budget constraint $x + z \leq I$. Again, the units of the goods are chosen so that both prices are 1, per-period income I is large compared to the consumption of x, and individuals do not save or borrow. When individuals maximize $\hat{u}(x, z)$ subject to the budget constraint, they allocate their income in a way that maximizes decision utility and is depicted by (x*, z*). The actual consumption of the unhealthy good satisfies $v_x(x^*; \alpha_\rho \rho) - \beta c_x(x^*; \alpha_\gamma \gamma) - 1 = 0$, and the consumption of the composite good is $z^* = I - x^*$. It is straightforward to see that self-control problems ($\beta < 1$) can lead to overconsumption of the sin good ($x^* > x^{**}$) (see O'Donoghue and Rabin, 2006). But also can distorted perceptions lead to deviations from optimal consumption. However, it is not a priori clear whether distorted perceptions lead to more or less consumption of the unhealthy good. Perceived costs and benefits of unhealthy consumption can be either higher or lower than actual costs and benefits. Note that different deviations from rationality can, in principle, cancel each other out.

4.2. Behavioral Policy Interventions

O'Donoghue and Rabin (2006) investigate the extent to which taxes can reduce differences between decision utility and experienced utility arising from present-biased preferences. We use the model to illustrate some effects of soft policy interventions. Assume that individuals have self-control problems ($\beta < 1$), but no distorted perceptions ($\alpha_\rho = 1$, and $\alpha_\gamma = 1$). A social planner realizes that individuals overconsume the unhealthy good. The planner decides to use behavioral insights to nudge individuals towards the choices that make them presumably better off, i.e. towards consuming x** instead of x*. In this model, there are three behavioral strategies that the social planner can use to achieve behavioral change: they can (a) make unhealthy consumption appear less favorable

by reducing α_ρ, they can (b) increase the perceived future costs of consuming unhealthy today by increasing α_γ, and they can (c) try to reduce the tendency to pursue immediate gratification and bring β to unity. We assume that behavioral interventions can change α_ρ, α_γ, and β for longer than only the current period.

First, assume that the policy-maker aims to reduce α_ρ to α_ρ^n ($\alpha_\rho^n < \alpha_\rho$), i.e. the policy-maker reduces the perceived benefits of the sin good (for example by using framing effects to reduce the attractiveness of junk food). Assume that the intervention is successful and the consumption of the unhealthy good is reduced so that $x^n = x^{**}$. Note, however, that now a situation has emerged, where two deviations from rationality cancel each other out. Consumption of the unhealthy good now satisfies $v_x(x^{**}; \alpha_\rho^n\rho) - \beta c_x(x^{**}; \alpha_\gamma\gamma) - 1 = 0$, so that the outcome is *as if* the underlying behavior was rational. However, since $\beta < 1$ and $\alpha_\rho^n < 1$, this is not the case.

Based on the rules of libertarian paternalism everything is fine as well-being has increased without decreasing freedom of choice.[28] A social planner engaging in AEP, however, would not implement this intervention because it does not help individuals to make well-informed, critically reflected decisions. To the contrary, the intervention introduces a new bias ($\alpha_\rho^n < 1$), which might become stronger over time and induce unconscious preference learning. Moreover, the new bias hides the present-bias from individuals' views. If interventions hide biases, individuals will be less likely to realize similarly biased behavior in future situations.[29]

The problematic character of behavioral interventions that do not enhance autonomy becomes particularly obvious when, for any reasons, the decision context changes over time (individuals might, for example, change the school and go to another cafeteria). In the first decision context, say at time t+0, individuals did not have the opportunity to understand the actual reasons for the overconsumption of the unhealthy good (actually there was no overconsumption). The individuals' present-biases are not reduced and β is still below unity. Moreover, as a result of the behavioral inter-

[28] Jolls and Sunstein (2006) explicitly refer to such strategies when discussing that some forms of bounded rationality can counteract other forms of bounded rationality. For example, they suggest that loss aversion can be neutralized by framing situations in a way that exploits individuals' optimism bias.

[29] Similarly, in a situation where two or more biases offset each other so that the behavior is nearly optimal, a social planner engaging in AEP would encourage correcting only one bias even if this meant a reduction of welfare in the short run (see Besharov, 2004).

vention at time t+0, the individuals might unlearn the capacity to engage in effective self-regulation. Self-regulation can be improved through regular exercise (Baumeister et al., 2006), and the described intervention may remove such opportunities. The intervention in period t+0 may thus even increase the present-bias over time so that β is reduced to β' ($\beta' < \beta$). Assume that in a new decision context at time t+1, no behaviorally informed policies are present anymore that could utilize the individuals' distorted perceptions to reduce unhealthy consumption. In this situation, the consumption of the unhealthy good satisfies $v_x(x'; \alpha_\rho\rho) - \beta' c_x(x'; \alpha_\gamma\gamma) - 1 = 0$. Hence, the consumption of the unhealthy good in the new decision-making context might be even higher than without having been exposed to the behaviorally informed policy in the first place (i.e. $x' > x^*$).[30] This example illustrates that when acknowledging dynamic effects, even well-motivated behavioral interventions can be detrimental for individuals' well-being. This is, however, only true when the interventions try to change the outcomes of decision-making without influencing the reasons for decision-making anomalies. This type of intervention is similar to curing the symptoms of a sickness without looking at the underlying reasons of the sickness.

Alternatively, assume that the policy-maker engages in AEP and tries to reduce individuals' tendency to impulsively pursue immediate gratification and thereby bring β to unity. To do so the policy-maker can try to (i) reduce the present-bias in System 1, and (ii) activate System 2. To reduce the present-bias the policy-maker can, for example, encourage individuals to choose smaller plates (Wansink and Cheney, 2005), and to adopt an abstract mindset or imagine tempting stimuli in a non-consummatory fashion, which both reduces temptations (Hofmann et al., 2010). The policy-maker can also make individuals aware of their present-bias and provide an understanding of the bias' origins. The policy-maker can transfer the knowledge about factors that can lead to impulsive behavior, such as fatigue (Baumeister et al., 1998), cognitive load (Shiv and Fedorikhin, 1999), visceral states such as hunger and thirst (Loewenstein, 1996), and being exposed to many attractive cues (Lades, 2012). Increased awareness and understanding of these factors allow individuals to effectively modify their own decision-making contexts and thus to engage in self-nudging (see Lades, 2014). To activate System 2 and encourage deliberative decision-making policy-makers can, for example, prompt individuals to make decisions well in advance (Rogers and Bazerman, 2008), provide pre-commitment mechanisms (Bryan et al., 2010), induce a higher construal level

[30] Whether behavioral interventions can increase decision-making anomalies over time, however, is essentially an empirical question worth pursuing in future research (see also de Haan and Linde, 2012).

thinking rather than lower level thinking (Trope and Liberman, 2003), and make people accountable for their decisions.[31]

V. Conclusion

Individuals can make mistakes and these can turn out to be welfare-decreasing for them. While not all departures from rational choice theory might necessarily mean lower well-being (e.g., Schwartz et al., 2002), there are enough instances of systematic non-optimal behavior. In this respect, behavioral economists are right in claiming that there is scope for paternalism (however defined). In the present paper, we have argued that soft paternalism, as exemplified by the idea of libertarian paternalism, has a major shortcoming in that it neglects the fact that behavioral interventions may reduce the individual capability to make critically reflected, autonomous decisions and that this becomes particularly pressing when considering the dynamic effects of behavioral interventions.

To remedy this shortcoming the paper has sketched a notion of "autonomy-enhancing paternalism" (AEP). In line with Dworkin (1988) and in the language of Kahneman (2011), we define autonomy as the possibility to make critically reflected decisions in System 2. Behavioral policy interventions that are in line with AEP aim to increase individual welfare over time by changing the choice architecture to facilitate individuals' ability to make critically reflected decisions. Interventions that enhance autonomy, but do not increase welfare cannot be justified within AEP as these interventions are not paternalistic. Also paternalistic interventions that have positive short-run effects on individual well-being, but reduce autonomy cannot be justified within AEP. Making people better off at the expense of their autonomy (especially through manipulative nudges of which people are not aware) risks "infantilizing" consumers (Barber, 2008) and is morally undesirable. While libertarian paternalism seems to approve of interventions that induce individuals to do the right thing for the wrong reasons, AEP insists on interventions that help individuals doing the right thing for the right (unbiased) reasons. AEP thus takes seriously the idea of sovereign citizens, renders manipulation impossible, accounts for the heterogeneity of individuals, and counteracts slippery slope arguments by decreasing the probability of future paternalistic interventions. Most importantly,

[31] Not all of these ways of reducing present-bias change the context in which decisions are made. Some of the mentioned interventions aim to educate the individuals, train them, and generate awareness and understanding (Soll et al., 2013).

AEP highlights the importance of autonomy in a dynamic framework, where policy interventions can shape preferences and the strength of decision anomalies over time. Interventions that do not enhance, or even reduce, individual autonomy can increase individuals' cognitive biases and induce unwanted preference learning paths (Binder, 2014). In our view, soft paternalism should modify decision contexts to help individuals to overcome their biases and decision-making fallibilities over time and thus encourage them to make better, autonomous choices.

Acknowledgments

Martin Binder was funded by the ESRC-TSB-BIS-NESTA as part of the ES/J008427/1 grant on Skills, Knowledge, Innovation, Policy and Practice (SKIPPY). We wish to thank the participants of the HEIRs conference 2013 on Public Happiness in Rome, of the SABE/IAREP/ICABEEP 2013 Conference in Atlanta, GA, of a seminar at the Strategic Organization Design unit in Odense, the workshop on "Global Economic Ethics" in Kassel 2014, and of the 2013 Evolutionary Economics seminar in Jena, as well participants of research seminars in Kassel and Berlin for helpful comments. In particular we like to thank Thomas de Haan, Robert Sugden, and Ulrich Witt. If behavioral economists are right, errors are bound to remain in this manuscript, of which we are not aware, but which - as usual - are solely our responsibility.

References

Ariely, D. & Norton, M. I. (2008). How actions create–not just reveal–preferences. *Trends in Cognitive Sciences*, 12(1):13–16.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: is the active self a limited resource?. *Journal of Personality and Social Psychology*, 74(5), 1252.

Baumeister, R. F., Gailliot, M., DeWall, C. N., & Oaten, M. (2006). Self-Regulation and Personality: How Interventions Increase Regulatory Success, and How Depletion Moderates the Effects of Traits on Behavior. *Journal of Personality*, 74(6), 1773-1802.

Barber, B. S. (2008). *Consumed - How Markets Corrupt Children, Infantilize Adults, and Swallow*

Citizens Whole. W.W.Norton & Company.

Berg, N., Eckel, C., & Johnson, C. (2011). Inconsistency pays?: Time-inconsistent subjects and EU violators earn more. Mimeo.

Berridge, K. C., & Aldridge, J. W. (2008). Decision utility, the brain, and pursuit of hedonic goals. *Social Cognition*, 26(5), 621.

Besharov, G. (2004). Second-best considerations in correcting cognitive biases. *Southern Economic Journal*, 71(1), 12-20.

Binder, M. (2010). *Elements of an Evolutionary Theory of Welfare*. Routledge, London.

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

Bovens, L. (2009). The ethics of Nudge. In: Grüne-Yanoff T, Hansson SO (eds) Preference change: approaches from philosophy, economics and psychology. Springer, Berlin, pp 207–220

Bryan, G., Karlan, D., & Nelson, S. (2010). Commitment devices. *Annual Review of Economics*, *2*(1), 671-698.

Camerer, C. (2004). *Advances in behavioral economics*. Princeton University Press.

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

Carlin, B. I., Gervais, S., & Manso, G. (2013). Libertarian Paternalism, Information Production, and Financial Decision Making. *Review of Financial Studies*, 26(9), 2204-2228.

Conlisk, J. (1996). Why bounded rationality? *Journal of Economic Literature*, 34(2):669–700.

de Haan, T., & Linde, J. (2012). Nudge lullaby. Working Paper.

Dworkin, G. (1988). *The Theory and Practice of Autonomy*. Cambridge University Press, Cambridge/UK.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223-241.

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press, New York/Oxford.

Gigerenzer, G. (2011). What are natural frequencies? BMJ; 343:d6386.

Gilbert, D. T. & Ebert, J. E. J. (2002). Decisions and revisions: The affective forecasting of changeable outcomes. *Journal of Personality and Social Psychology*, 82(4):503–514.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. (2010). Ego depletion and the strength model of self-control: a meta-analysis. *Psychological Bulletin*, 136(4), 495.

Hanks, A. S., Just, D. R., & Wansink, B. (2013). Smarter lunchrooms can address new school lunchroom guidelines and childhood obesity. *The Journal of Pediatrics*, 162(4), 867-869.

Hausman, D. M. & Welch, B. (2010). Debate: To nudge or not to nudge. *Journal of Political Philosophy*, 18(1):123–136.

Hergenhahn, B. R. & Olson, M. H. (1997). *An Introduction to Theories of Learning*. Prentice Hall, Upper Saddle River/New Jersey, 5th edition.

Hofmann, W., Deutsch, R., Lancaster, K., & Banaji, M. R. (2010). Cooling the heat of temptation: Mental self-ontrol and the automatic evaluation of tempting stimuli. *European Journal of Social Psychology*, *40*(1), 17-25.

Iyengar, S. S., Jiang, W., & Huberman, G. (2003). How much choice is too much?: Contributions to 401(k) retirement plans. Mimeo.

Jolls, C., & Sunstein, C. R. (2006). Debiasing through law. *The Journal of Legal Studies*, 35, 1.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics*, 375-405.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93(5):1449–1475.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar Straus & Giroux.

Korobkin, R. (2011). What comes after victory for behavioral law and economics. *Unversity of Illinois Law Review*, 2011(5):1653–1674.

Lades, L. K. (2012). Towards an incentive salience model of intertemporal choice. *Journal of Economic Psychology*, 33:833–841.

Lades, L. K. (2014). Impulsive consumption and reflexive thought: Nudging ethical consumer behavior. *Journal of Economic Psychology*, 41, 114-128.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–477.

Larrick, R. P. (2004). *Debiasing*. In D. J. Koehler & N. Harvey (Eds.), Blackwell Handbook of Judgment and Decision Making (pp. 316-337). Oxford: UK.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65(3), 272-292.

Loewenstein, G. & Haisley, E. (2008). The economist as therapist: Methodological ramifications of 'light' paternalism. To appear in A. Caplin and A. Schotter (Eds.), "Perspectives on the Future of Economics: Positive and Normative Foundations", volume 1 in the Handbook of Economic Methodologies, Oxford, England: Oxford University Press.

Mills, C. (2013). Why nudges matter: A reply to Goodwin. *Politics*, 33(1), 28-36.

Mullainathan, S., Schwartzstein, J., & Congdon, W. J. (2012). A reduced-form approach to behavioral public finance. *Annual Review of Economics*, *4*(1), 511-540.

O'Donoghue, T. & Rabin, M. (2003). Studying optimal paternalism, illustrated by a model of sin taxes. *The American Economic Review*, 93(2):186–191.

O'Donoghue, T. & Rabin, M. (2006). Optimal sin taxes. *Journal of Public Economics*, 90(10):1825–1849.

Qizilbash, M. (2012). Informed desire and the ambitions of libertarian paternalism. *Social Choice and Welfare*, 38(4):647–658.

Rabin, M. (1998). Psychology and economics. *Journal of Economic Literature*, 36(1):11–46.

Rebonato, R. (2012). *Taking Liberties - A Critical Examination of Libertarian Paternalism*. Palgrave- Macmillan, Basingstoke.

Rizzo, M. J. & Whitman, D. G. (2009). Little brother is watching you: New paternalism on the slippery slopes. *Arizona Law Review*, 51(3):685–739.

Rogers, T., & Bazerman, M. H. (2008). Future lock-in: Future implementation increases selection of 'should'choices. *Organizational Behavior and Human Decision Processes*, *106*(1), 1-20.

Schubert, C. (2012). Pursuing happiness. *Kyklos*, 65(2):245–261.

Schubert, C. & Cordes, C. (2013). Role models that make you unhappy: Light paternalism, social learning and welfare. *Journal of Institutional Economics*, 9(2):131–159.

Schwartz, B. (2000). Self-determination - the tyranny of freedom. *American Psychologist*, 55(1):79–88.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology*, 83(5):1178–1197.

Shiv, B., & Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research*, *26*(3), 278-292.

Soll, J. B., Milkman. K. L., & Payne, J. W. (2013). A user's guide to debiasing. Working Paper.

Sunstein, C. R. (2013). *Simpler: The future of government*. Simon and Schuster.

Sunstein, C. R. (2014). *Why Nudge?: The Politics of Libertarian Paternalism*. Yale University Press.

Sunstein, C. R. & Thaler, R. H. (2003). Liberaterian paternalism is not an oxymoron. *The University of Chicago Law Review*, 70(4):1159–1202.

Sunstein, C. R., & Reisch, L. A. (2013). Green by default. *Kyklos*, *66*(3), 398-402.

Thaler, R. & Sunstein, C. R. (2003). Libertarian paternalism. *The American Economic Review*, 93(2):175– 179.

Thaler, R. H. & Sunstein, C. R. (2008). *Nudge - Improving Decisions about Health, Wealth and Happiness*. Penguin Books, London.

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review*, *110*(3), 403.

Trout, J. D. (2005). Paternalism and cognitive bias. *Law and Philosophy*, 24:393–434.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Wansink, B., & Cheney, M. M. (2005). Super bowls: serving bowl size and food consumption. *JAMA: the Journal of the American Medical Association*, *293*(14), 1727-1728.

Witt, U. (1991). Economics, sociobiology, and behavioral psychology on preferences. *Journal of Economic Psychology*, 12:557–573.

Witt, U. (2001). Learning to consume - a theory of wants and the growth of demand. *Journal of Evolutionary Economics*, 11:23–36.

Witt, U. & Binder, M. (2013). Disentangling motivational and experiential aspects of "utility" - a neuroeconomics perspective. *Journal of Economic Psychology*, 36(1):27–40.

Zajonc, R. B. & Markus, H. (1982). Affective and cognitive factors in preferences. *Journal of Consumer Research*, 9(2):123–131.