PAPERS on Economics & Evolution

1304

Autonomy-enhancing paternalism

by

Martin Binder
Leonhard K. Lades

The *Papers on Economics and Evolution* are edited by the Evolutionary Economics Group, MPI Jena. For editorial correspondence, please contact: evopapers@econ.mpg.de

ISSN 1430-4716

Max Planck Institute of Economics
Evolutionary Economics Group
Kahlaische Str. 10
07745 Jena, Germany
Fax: ++49-3641-686868

Autonomy-enhancing paternalism

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

[a]*Universität Kassel, Fachbereich Wirtschaftswissenschaften, Lehrstuhl für Wirtschafts- und Unternehmensethik, Nora-Platiel-Str.4, 34109 Kassel, Germany*

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

[c]*Science and Technology Policy Research Unit, University of Sussex, Falmer, Brighton, BN1 9SL, UK*

(binder@uni-kassel.de, lades@econ.mpg.de).

Abstract

Behavioral economics has shown that individuals sometimes make decisions that are not in their best interest. This insight has prompted calls for behaviorally-informed policy interventions popularized under the notion of "libertarian paternalism". This type of soft paternalism aims at helping individuals without reducing their freedom of choice. We highlight three problems of libertarian paternalism: the difficulty to detect what is in the best interest of an individual, the focus on freedom of choice at the expense of a focus on autonomy, and the neglect of the dynamic effects of libertarian paternalistic policy interventions. We present a form of soft paternalism called "autonomy-enhancing paternalism" that seeks to constructively remedy these problems. Autonomy-enhancing paternalism suggests using insights from subjective well-being research in order to determine what makes individuals better off. It imposes an additional constraint on the set of permissible interventions highlighting the importance of autonomy in the sense of the capability to make critically reflected, i.e. autonomous, decisions. Finally, it acknowledges that behavioral interventions can change the strength of individual decision making anomalies over time as well as influence individual preference learning. We illustrate the differences between libertarian paternalism and autonomy-enhancing paternalism in a simple formal model in the context of optimal sin nudges.

Keywords: libertarian paternalism, behavioral economics, subjective well-being, autonomy, preference learning, welfare economics

1. Introduction

Behavioral economics has shown that individual decision-making is not always characterized by full rationality, perfect information, 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 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 normative implications of the insights of behavioral economics are known under the label of "*libertarian paternalism"* (Thaler and Sunstein, 2003; Sunstein and Thaler, 2003; Thaler and Sunstein, 2008). Libertarian paternalists argue that the deviations from standard economic rationality prompt for the benign intervention by a social planner (in the phrase 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" (Thaler and Sunstein, 2008) 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 significantly making alternatives more costly. The debate sparked by libertarian paternalism provides a useful stimulus to explore the normative and political consequences of behavioral economics. Correspondingly, libertarian paternalism has found a number of prominent defenders (e.g., Camerer, 2006; Trout, 2005; Loewenstein and Haisley, 2008; Amir and Lobel, 2008; Desai, 2011; Sunstein, 2012) and an equal if not greater number of critics (Sugden, 2008; Rizzo and Whitman, 2009a,b; Binder, 2013).

In this paper, we focus on three important, interconnected, and mutually dependent problems of libertarian paternalism. These problems are (1) the difficulty for the choice architect to detect what is in the best interest of an individual, (2) the focus on "freedom of choice" at the expense

of a focus on "autonomy" in the sense of critical reflection of decisions,[1] and (3) the neglect of the dynamic effects of libertarian paternalistic policy interventions, which usually are analyzed in a timeless one-shot perspective (see Binder, 2013, on these points). In order to overcome these three shortcomings we present a form of soft paternalism, which we call "autonomy-enhancing paternalism" (AEP), which adopts attractive insights of libertarian paternalism, but avoids its most serious shortcomings.

The paper is structured as follows. Section 2 briefly summarizes the concept of libertarian paternalism. Section 3 elaborates on the three mentioned weaknesses of libertarian paternalism and suggests "autonomy-enhancing paternalism" (AEP) as a remedy for the shortcomings. AEP has three crucial elements: It suggests using well-founded insights from subjective well-being research in order to determine what makes individuals better off in a substantive and systematic way. Taking the notion of autonomy seriously (as also suggested, e.g., by Hausman and Welch, 2010), it prevents that behavioral policy interventions manipulate individuals by exploiting their decision-making anomalies. Instead, it encourages individuals to make critically reflected, autonomous decisions. Finally, AEP takes a decidedly dynamic perspective and acknowledges that behavioral interventions can and typically will change the strength of individual decision making anomalies as well as influence individual preference learning. Thereby AEP provides new and important arguments that should be considered when engaging in cost-benefit analyses of behavioral policy interventions. Section 4 illustrates the differences between libertarian paternalism and autonomy-enhancing paternalism in a simple formal model in the context of optimal sin nudges. Section 5 concludes.

2. Libertarian paternalism

Traditionally, economists tend to be wary of paternalism and "robust[ly]" (Sugden, 2008, p. 226) oppose it justifying their position by relying on "epistemic privilege" (Buchanan, 1991), i.e. the

[1] We define autonomy in the sense of being able to make a critically reflected decision (Dworkin, 1988). We will elaborate on this idea of autonomy in a later section. Other notions of autonomy (see, e.g., Christman, 2011) might not apply to this paper's reasoning.

notion that the individual knows best of its own interests and no social planner could know better. While the idea that the individual knows best about its interests indeed makes a case for non-interference, it is this very idea that has come under attack through behavioral economics. Mainly when decisions are made by the automatic, intuitive "System 1", libertarian paternalists emphasize "the possibility that in some cases individuals make inferior choices, choices that they would change if they had complete information, unlimited cognitive abilities, and no lack of willpower" (Thaler and Sunstein, 2003, p. 175).[2] These instances of bounded rationality then justify the exploration of behavioral interventions, which are called "nudges". The first main difference to traditional paternalism is that while traditional paternalism interferes with all types of decision-making, libertarian paternalism suggests interfering only with those decisions made in the intuitive System 1. The behavior of individuals who deliberatively and cognitively reflect upon their decisions in System 2 is supposed not to change as a result of the libertarian paternalistic nudges.

Based on this reasoning, libertarian paternalists aim at influencing individuals' choices "in a way that makes choosers better off, as *judged by themselves*'' (emphasis in the original, Thaler and Sunstein, 2008, p. 5). The aim of making the nudged individuals better off by interfering with their choices classifies the approach as being *paternalistic*. To detect the decisions that the individuals judge to be in their best interests, libertarian paternalists rely on conjectures about the behavior that is not influenced by any decision anomaly identified in behavioral economics. Hence, libertarian paternalistic interventions aim at steering choices to those outcomes that the nudged individuals would have chosen for themselves if they were not influenced by the decision anomaly in question.

The second distinction between *libertarian* paternalism and traditional (hard) paternalism is to design the interventions so as to leave individual freedom of choice intact. The emphasis "is not on blocking choices, but on strategies that move people in welfare-promoting directions while

[2] Kahneman (2011) popularized the notions of "System 1" and "System 2" to distinguish two ways of thinking and making decisions. While "System 1" thinking is fast, uncontrolled, unconscious, intuitive, and affective, "System 2" thinking is reflective, slow, deductive, and cognitive. To be fair, one could also argue that some System 2 choices are not in the individual's best interest and hence legitimize paternalism. The literature is divided on the exact criterion to demarcate choices into being in an individual's interests or not.

also allowing freedom of choice" (Sunstein and Thaler, 2003, p. 1170). Moreover, only when economic incentives are not changed significantly, a behavioral intervention can be considered as being libertarian (Hausman and Welch, 2010, p. 126). Libertarian paternalistic interventions are supposed to merely nudge individuals towards certain choices. Individuals shall be able to reverse the nudge at the lowest possible cost if they do not agree with the choice architect if engaging in critical reflection in the decision-making System 2.

Perhaps the most (in)famous example for a libertarian paternalistic policy intervention is the arrangement of food in a school cafeteria, so as to more prominently place healthy foods (Thaler and Sunstein, 2008). While this arrangement can encourage healthy food choices, the arrangement does not reduce freedom of choice or significantly change incentives (unhealthy foods are available but somewhat out of sight). Another famous example is the setting of default rules, such as automatic enrollment (with the option to opt out) to pension schemes (Choi et al., 2004; Madrian and Shea, 2001). Ill-informed or weak-willed individuals are nudged towards welfare-improving behavior (saving for their pensions), while they are still allowed to opt out with no or just small costs (such as writing a cancellation letter). Other instruments in the libertarian paternalistic toolkit are "required active choosing" (mandated choice), "cooling off periods" (individuals can impulsively purchase things at will, but are granted some time to reverse their decisions), and the "debiasing" and "reframing" of consumer choices (Trout, 2005).

Libertarian paternalism has sparked a lively debate in recent years, which has offered scholars the opportunity to explore the normative implications of behavioral economics. A typical argument in its favor holds that large welfare gains can be realized at low costs by helping less thoughtful individuals to make decisions that make them better off. Compared to these welfare gains, reductions in freedom of choice are argued to be small (e.g., Trout, 2005; Thaler and Sunstein, 2008; Camerer et al., 2003). Especially in situations that are complex and where people lack experience in decision-making, libertarian paternalism tends to look plausible (Loewenstein and Haisley, 2008, p. 8). Moreover, choice architecture is often inevitable. Objections against libertarian paternalism come from a variety of directions, most viciously from libertarians who flat out deny that libertarian paternalism is libertarian or in any way liberty-preserving (Mitchell, 2005; White,

2008). More traditional critiques in a public choice theoretic spirit doubt the will and ability of government representatives to correct bias (Glaeser, 2005). But libertarian paternalism could also be attacked from the other side of the political spectrum for not being intrusive enough to solve the most pressing societal problems. We do not wish to survey the full spectrum of arguments pro and contra libertarian paternalism (see Rebonato, 2012, for a rather comprehensive overview), but rather focus on three important shortcomings of libertarian paternalism, which we hope to remedy with our proposal for "autonomy-enhancing paternalism".

3. Autonomy-enhancing paternalism

3.1. How to identify individual judgments about what is in one's best interest?

Our first objection to libertarian paternalism rests on the question whether the choice architect can reliably detect the choices that make individuals better off as judged by themselves. In the neoclassical framework, this is not an issue. In this framework, an individual's choice for, say, an apple instead of a banana reveals that the apple makes the individual better off. Choices reveal true preferences by assumption (Samuelson, 1938). Moreover, individuals' predictions about which choices make them better off cohere perfectly with their choices.[3] In the neoclassical framework, policy-makers can thus easily identify what makes individuals better off as judged by themselves by simply observing individual choices. Behavioral economics departs from these assumptions and focuses on more realistic cases where individuals' predictions about what makes them better off can be mistaken (see Wilson and Gilbert, 2005; Witt and Binder, 2013) and where visceral factors can influence individual choices in non-optimal ways (Loewenstein, 1996; Lades, 2012). In these cases, neither predictions, nor choices reveal what actually makes individuals better off as judged by themselves. Individuals can err and (plan to) choose the apple although choosing the banana would have turned out to increase welfare. The choice architect thus lacks an appropriate standard to assess what makes individuals better off as judged by themselves.

[3] In the utility taxonomy of Kahneman et al. (1997), the neoclassical framework can be described as follows: Individual judgments about what makes them better off ("predicted utility") directly lead to individual choices ("decision utility"), which by definition make the individuals better off ("experienced utility").

To illustrate assume that an individual endowed with an income of *I* has the choice between two products x_A and x_B with prices p_A and p_B, respectively. Given that the individual rationally maximizes her utility, she will solve

$$\max U(x_A,x_B) \quad \text{s.t.} \quad p_Ax_A + p_Bx_B \leq I \qquad (1)$$

and obtain the optimal levels of consumption. Since the individual is rational, a choice of x_A reveals that the individual's preference for x_A is stronger than that for x_B. No external knowledge is needed about what x_A and x_B represent, whether the individual indeed predicts x_A to be better than x_B, or whether indeed x_A makes the individual better off than x_B. But if we allow that individuals sometimes misjudge what makes them better off and (also because of that) deviate from utility maximizing behavior, the description of the decision-making process in equation 1 is no longer accurate. Let's describe a deviation from economic rationality by an error term ε, which might occur when predictions and/or decisions are made intuitively (we subtract this error term to denote the shortfall in rationality).[4] Integrating this error term into equation 1, the above decision can be described by

$$\max\text{-}\varepsilon\ U(x_A,x_B) \qquad \text{s.t.} \quad p_Ax_A + p_Bx_B \leq I \qquad (2)$$

In this admittedly simplified "behavioral economic" description of individual decision-making, System 1 is represented by ε, and System 2 is represented by those parts of equation 2 that are similar to equation 1 (see Kahneman, 2011).[5] In this model, a choice for x_A does not necessarily reveal that x_A makes the individual better off compared to x_B. The choice for x_A can also be triggered by a mistake in the prediction about what makes the individual better off or by a decision-making error, both reflected by ε.

Some behavioral economists have argued that even when individuals deviate from utility maximizing behavior, it is possible to identify individual preferences in the sense of what makes individ-

[4] 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.)

[5] Note that while errors tend to occur more frequently in System 1 than in System 2, most proponents of 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.

uals better off (Beshears et al., 2008; Köszegi and Rabin, 2008).[6] One prominent way to do so is to rely on "informed" preferences. Proponents of libertarian paternalism seem to refer to these informed preferences when, for example, speaking about inferior choices that individuals "would change if they had complete information, unlimited cognitive abilities, and no lack of willpower" (Thaler and Sunstein, 2003, p. 175). In context of green defaults, for example, Sunstein and Reisch (2013) argue that the "preferred approach is *to select the default rule that reflects what most people would choose if they were adequately informed*" (p. 401, italics in the original). The question arises whether the concept of informed preferences can help the choice architect to use choice data in order to identify what is in the individuals' best interests as judged by themselves.[7] We believe that this is very difficult if not impossible. It is not obvious how the choice architect can detect whether a choice is "informed" or not. Since no one has unlimited cognitive abilities and no lack of willpower, one has to relax these requirements.[8] When doing so, it is unclear how to operationalize the point at which a preference is informed enough. How much willpower, how much self-control is needed for a preference to count as informed (see Camerer et al., 2003, p. 1214)? These questions are often left unanswered or if they are answered, they are answered with widely different definitions of when a preference should count as informed. Moreover, to make an individual's hypothetical set of informed preferences the measuring rod for rational behavior may be so demanding that failing to conform to those can hardly be called a fault of human decision-making (but rather a fault of these unattainable standards; see Sugden, 2011, pp. 24-25).[9] Utilizing informed preferences as the measuring rod for welfare therefore gives legislators a ready excuse to interfere with individual decisions whenever the individuals do not rise to

[6] Beshears et al. (2008) discuss six approaches that use behavior to identify normative preferences while acknowledging that revealed preferences often differ from normative preferences: structural estimation, active decisions, asymptotic choice, aggregated revealed preferences, reported preferences, and informed preferences. They argue that policy interventions should be informed by insights from all six approaches.

[7] A second question arises whether, if at all possible, it is good to base policy implications of individuals' judgments about what makes them better off (predicted utility) rather than on what actually makes the individuals better off (experienced utility). We will come back to this question.

[8] Given this definition, some researchers even argue that informed preferences are so different from individuals' actual preferences that it is not clear whether the informed individual would be the same individual as the individual having its actual preferences (see, e.g., Sobel, 1994; Qizilbash, 2012).

[9] While libertarian paternalists describe individual behavior as falling short of perfect rationality, their normative view is still to demand that individual behavior should accord to perfect economic rationality. In this, we need to acknowledge that behavioral economics in general still lacks an appropriate standard of normative rationality.

the Olympian heights of rationality, i.e. pretty much always. Setting an unattainable standard of rationality thus amounts to giving a blanket justification for paternalistic intervention (Qizilbash, 2011, p. 37). Informed preference views have been successfully refuted in normative economics for quite some time but, not unlike a zombie, refuse to die (Sobel, 1994; Rosati, 1995; Qizilbash, 2012).

Another way of identifying individual's judgments about what makes them better off using choice data when acknowledging that individuals can make mistakes is to rely on choices made by the decision-making System 2 (Kahneman, 2011). Thaler and Sunstein (2008), Ch. 2, suggest using a variant of this the distinction between System 1 and System 2 (fallible "humans" versus perfectly rational "econs") and privilege the preferences of "econs". However, it is not simple to determine whether a decision is made exclusively by System 2 and is not affected by System 1. How small should ε be so that preferences reveal individuals' judgments about what makes them better off? Moreover, even System 2 choices can fall short in improving our welfare, and some researchers question whether deviations from rational choice are actually detrimental to human well-being (Berg et al., 2011a,b). Most critically, System 2 choices are only defined procedurally but not substantively. The choice architect is thus still left with hypothesizing which are the substantive "true" System 2 preferences an individual would hold (leading to the associated problems known from informed preference views).

As an alternative to using revealed preferences to detect individuals' judgments about what makes them better off, Loewenstein and Haisley (2008) argue in favor of a "pragmatic approach" that aims at flexibly using a wide range of information about individuals' interests. Thaler and Sunstein's (2008) "New Year's resolution test" (p. 73) is similar to this pragmatic approach, by recurring to some sort of reflected metapreference declared as "New Year's resolution". But in essence, these pragmatic definitions of what individuals judge to be in their best interests open the door to ad hoc standards of what paternalists judge to be welfare promoting for the individuals. The appeal to a pragmatic approach is prone to arbitrary value judgments by the choice architect (see more extensively Binder, 2013). Moreover, these types of tests, which, in essence, could be seen as stated (meta-)preferences are influenced by "faulty affective forecasting" and related

mispredictions of future well-being, where individuals focus on salient aspects of a decision or neglect hedonic adaptation when predicting their future well-being and forming preferences based on this (Wilson and Gilbert, 2005; Gilbert, 2007). Stated (meta-)preferences can thus be as fallible as revealed preferences and distortions in judgments about future utility cast doubt on the idea that such statements will indeed reflect what makes the individual better off in the future.

In our view, the main problem of these approaches to detect what makes individuals better off as judged by themselves is that they do not deal with the content of actual choices. These approaches exclusively deal with the decision making mechanisms that underlie choices for unspecified goods. What makes individuals better off is only defined procedurally, i.e. types of decision-making are defined that shall reveal the individuals' judgments about what makes them better off. We argue that a substantive theory about which choices actually make individuals better off is needed. Equation 2, for example, indicates that the way in which individuals decide between x_A and x_B is prone to an error reflected by ε. However, the equation is silent about what x_A and x_B represent. We suggest that in order to detect what makes individuals better off it is beneficial to open the black-box of the utility function and to analyze which choices actually make individuals better off and why, but for this we need a substantive theory of the good (Binder, 2010, Witt and Binder, 2013). An additional problem in the libertarian paternalistic framework is that it is unclear whether "as judged by themselves" refers to the individuals' ex ante predictions about what makes them better off, or to what actually makes them better off ex post.

To open the black-box of the utility function one promising way is to look at subjective well-being research (Dolan et al., 2008; Easterlin, 2003; Layard, 2005; Kahneman et al., 1999, Binder and Coad, 2010, Welsch 2002). We suggest that taking recourse to subjective well-being (SWB) research offers a possible way to obtain an appropriate normative foundation on which behavioral policies can be based (Diener and Seligman, 2004, Binder, 2013). Subjective well-being can be used as an empirically well-founded basis for choice architects to determine what actually is in the individuals' best interests as judged by themselves (these measures are proven to be valid and reliable by psychological research, see, e.g. Krueger and Schkade, 2008). By using empirical

facts from subjective well-being research, clear statements can be given as to which policy interventions are (on average) in individuals' interests. Empirical research on the causes and correlates of SWB provides systematic, non-ad-hoc knowledge about individuals' likings, and a handful of determinants such as being healthy, enjoying an active social life, and being in employment are reliably found to explain large parts of well-being variation the world over (Graham, 2009, 2011). Based on this, and more concrete findings on the determinants of SWB, one can, for example, say that activity x_A makes individuals on average better off (happier) than activity x_B. Additionally, by using ex-post-SWB one avoids relying on statements possibly based on "faulty affective forecasting" (Wilson and Gilbert, 2005; Gilbert, 2007). It is thus better to base policy interventions on knowledge about which choices actually make individuals better off (ex post), rather than to base interventions on the individuals' (possibly wrong) predictions (ex ante). Here, behaviorally informed paternalism can be justified in a number of empirically well-established situations where it was shown that individuals make inferior choices that decrease their well-being in intertemporal settings even though their preferences are informed or based in the reflected System 2.

However, economists have only recently begun to engage in SWB research, and a note of caution is in order here before uncritically using insights from such research to determine which behavioral intervention makes individuals better off. A first complication stems from the fact that most subjective well-being research describes average effects of certain life events/changes on well-being, and hardly takes idiosyncrasies into account (but see Becchetti et al., 2008, Binder and Coad, 2011, Boyce and Wood, 2011, for applications taking into account the intricacies of subjective well-being and its correlates). More insights about inter-individual differences are needed to improve the ability of subjective well-being data to be used as a measure to define what makes individuals better off. A second complication lies in strategic responses of individuals to SWB surveys in order to influence policy interventions (Frey and Stutzer, 2010; this is mostly a problem for small specialized surveys where individuals can outguess the policy-maker's intention of the survey).

While we see potential for subjective well-being research as being able to inform an empirically well-defined standard for what makes individuals better off, we need to acknowledge the nascent

state the field is still in. But even if we had complete knowledge of the causes and correlates of human happiness, our point is that any proposal for soft paternalism should make room for individual autonomy.

3.2. The role of autonomy

Our first objection dealt with aspects of the "paternalistic" part of libertarian paternalism. We have argued that it is almost impossible for the choice architect to identify those individual judgments about what makes themselves better off that are not influenced by decision-making errors. Moreover, we argued that it may be better to base policy implications on insights about which choices actually make individuals better off, rather than on individual predictions. Our second objection deals with the "libertarian" aspects of the concept. In the libertarian paternalistic logic, policy interventions can be called libertarian when individual liberty is preserved and when incentives are not changed significantly. However, libertarian paternalists use a very narrow definition of what constitutes an individuals' liberty (namely: nominal freedom of choice; see on this line of critique e.g., Qizilbash, 2009).[10] Liberty, however, can also be understood in terms of the ability to make critically reflected, i.e. autonomous, decisions. Following Gerald Dworkin, we understand 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). Putting emphasis on the ability to make critically reflected, autonomous decisions is an important feature of liberal and libertarian views. However, this feature is not sufficiently taken into account by libertarian paternalists (Korobkin, 2011). Acceptable forms of (soft) paternalism should not exclude concerns about autonomy in favor of a narrow definition of liberty in terms of freedom of choice. (Real) Liberty can be only preserved if individuals can actually make autonomous decisions.

[10] 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 their 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).

To illustrate the importance of considering autonomy when evaluating behavioral policy interventions consider a benevolent planner who predicts that equation 1 is the best approximation for individual behavior. This planner aims at reducing the consumption of x_B depicting, say, junk food. Assume that the planner can choose between three different types of policy interventions: a ban, a tax, and a libertarian paternalistic nudge. Banning x_B will change equation 1 to

$$\max U(x_A) \qquad \text{s.t.} \qquad p_A x_A \leq I. \qquad (3)$$

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

$$\max U(x_A,x_B) \quad \text{s.t.} \quad p_A x_A + p_B^{tax} x_B \leq I, \qquad (4)$$

where $p_B^{tax} > p_B$. 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). Hence, they cannot be considered libertarian according to the libertarian paternalistic definition. Nudging individuals away from junk food, however, does not change equation 1. It is easy to think that in order to evaluate whether a policy intervention can be considered libertarian, libertarian paternalism utilizes the logic represented by equation 1.

But libertarian paternalism is informed by a behavioral economic model such as the one sketched in equation 2. It is this model that should be used to evaluate the acceptability of behavioral interventions. Equation 2 offers a way to incorporate our definition of autonomy, i.e. the possibility to make a critically reflected decision. Using the language of behavioral economic dual-process theories (e.g., Kahneman, 2011), one can understand autonomous decision making procedurally, viz. as being closely related to decision making in the reflective System 2. An autonomous person can be defined as somebody who has the possibility to let their reflective System 2 be responsible for the decision. This definition does not exclude the possibility that an autonomous person deliberatively decides to let ε influence some of their decisions. However, this autonomous person is always able to put System 2 back into power again. Autonomy is then the possibility to make a critically reflected decision in System 2 that is not hindered by any external or internal force.[11]

[11] This definition of autonomy focuses on internal forces that reduce the ability to make critically reflected decisions. Other authors have focused on external threats to autonomy (compare Christman, 2011).

When an individual, for example, cannot resist the temptation triggered by the perception of a chocolate cake, the decision to feast is not an autonomous one in this definition. 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. We use System 2 decision-making as a procedural restriction on behavioral policy interventions that has to be added as a safe guard against manipulation, while at the same time substantively defining individuals' interests with reference to their subjective well-being.

We suggest adding autonomy operationalized as a new constraint on the set of permissible behavioral interventions. Only those interventions and forms of nudges are legitimate that improve (or at least: leave intact) autonomy understood as the ability to let the reflective System 2 decide. Some nudges are thus compatible with the promotion of personal autonomy and can promote self-empowerment (see also Mills, 2013). In the best case, behavioral paternalistic interventions help foster critical thinking in System 2. At the minimum, interventions must not reduce individual autonomy in the sense of reducing the possibility to engage in critical thinking in System 2. Utilizing ε to change individual behavior can be considered manipulation even when the freedom to choose from the whole choice set is still nominally present. In our view, behavioral paternalistic interventions should not be manipulative even if the outcomes of the manipulation are in the interests of the manipulated individual. Manipulation is morally undesirable and contravenes the idea of individual autonomy and treating persons as ends in themselves. Manipulating individuals also 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).

While several nudges discussed in the literature already fulfill this acceptability condition, other nudges do not. Inacceptable nudges include those that harness bias and decision anomalies to "trick" behavior towards certain outcomes. Even when the choice architects conjecture (or even know with certainty) that these outcomes are in the best interest of the individual, the nudges are not acceptable when they reduce autonomy. Setting anchors that the individual does not know about would be not permissible. Also utilizing "yeah-whatever heuristics" (relying on individual inertia and laziness, see Thaler and Sunstein, 2008) and setting defaults without making

individuals aware of them would reduce autonomy. In particular behavioral policy interventions that aim at neutralizing one decision anomaly, say ε_1, by a nudge that utilizes another decision anomaly, say ε_2, constitute unacceptable nudging, if individuals are unawares of this.

On the other hand, acceptable nudges (i) reduce the effects of ε on individual decision-making, and (ii) activate System 2 decision-making (for an overview on debiasing, see Larrick, 2004). For example, presenting information in a simple and logical way that helps individuals to make well-informed decisions does increase autonomy since it reduces ε: Recently, "myplate" replaced the food pyramid in the U.S. as a much simpler way to illustrate the five food groups that are the building blocks for a healthy diet. Also, reducing choice sets to magnitudes that do not overwhelm individuals so that they fail to come to autonomous decisions (Trout, 2005; Camerer et al., 2003) is an acceptable nudge based on our acceptability condition. Providing all choice options may provide the maximal freedom of choice, but it is likely to reduce individual autonomy to make a critically reflected decision.[12] It may be autonomy-promoting to limit freedom of choice and thereby reduce ε. Another way to reduce the effects of ε is to make individuals aware of decision-making anomalies identified in behavioral economics. Other acceptable nudges activate System 2 so that although ε is present, it is not responsible for making decisions. For example, required active choosing mechanisms can bring to attention the possibility to decide in situations where this possibility would have not been obvious and probably neglected by the individual (Hausman and Welch, 2010, p. 134). Additionally, this helps individuals to recognize these decisions in future contexts (this is an application of the type of dynamic thinking we discuss below). We also consider cooling-off periods as acceptable, 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.

3.3. Dynamics

The third major problem of libertarian paternalism we wish to overcome is its mainly static out-

[12]Requiring a person to choose from thousands of mutual funds for their pension schemes has been empirically shown to decrease choice rates as opposed to choice situations with only few mutual funds (Iyengar et al., 2003).

look. Many libertarian paternalist proposals 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 nudges that have the ability to correct a given choice and thus increase individual well-being, even at the expense of autonomy. However, acknowledging that nudges often also influence future choices can change the evaluation of the nudges and needs to be explicitly considered by the choice architect. We highlight two interdependent ways how nudges can affect future choices: (a) via preference learning and (b) via changing the strength of the deviation from rationality, i.e. via changing ε.[13]

Regarding the nudges' 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 preferences (Ariely and Norton, 2008). Economic processes and policies, including soft paternalistic interventions, can shape preferences. Formally, a nudge at time t may (and likely will) influence the choice set in time t+1 and afterwards. For example, a nudge that increases the propensity to choose x_A may reduce the propensity to choose x_B. Nudges that do not reduce freedom of choice in the same period may nonetheless reduce freedom of choice in future periods. One should be aware of such effects when evaluating libertarian paternalistic policy interventions.

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). In this case, nudges are more effective but more problematic as well. 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.[14] If unconscious preference learning mechanisms

[13] Some critics also argue that libertarian paternalist interventions can increase the probability of further paternalist interventions, thus creating a slippery slope with regard to the number and types of interventions (e.g., Rizzo and Whitman, 2009b). With the abandoning of a clear-cut rationality measure (see our first objection), one can imagine how a once drawn line is progressively shifted over time. In this paper, however, we emphasize dynamic changes occurring within the nudged individuals, not within policy-makers or societal norms.

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

are (inadvertently) exploited by libertarian paternalists, 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).[15] 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). Even if, at one point in time, individuals were to become aware of the nudge's influence on their preference learning, the stickiness of associative preference learning can actually negate the possibility of making a critically-reflected decision in the future.[16]

Besides influencing preference learning, nudges can also influence individuals' abilities to learn from their mistakes and thereby influence ε over time. If individuals, for example, learn to rely on default rules set by the government or other choice architects (in this case ε refers to inertia and preferences for a status quo), ε can become stronger over time (Carlin et al., 2010; de Haan, 2011). Individuals may learn to be dependent and inactive. This creates the danger of a society where policy-makers are able to trick and manipulate. Behavioral policy interventions should always be transparent about both static and dynamic effects regarding welfare and autonomy.

Those nudges that help individuals to make better decisions over time are likely to increase autonomy by reducing ε, activating System 2, or making individuals aware of ε. Only when individuals know that ε may have an influence on their current and future behavior, they can critically reflect upon whether they are willing to let their behavior be influenced by ε. Simplifying decisions so that individuals are not overwhelmed by too big choice sets, for example, helps individuals to learn about their true preferences and thus can decrease ε over time. Many decision making situations can thus be understood as learning opportunities, where nudges can help individuals to make better informed and more autonomous choices in the future. This idea of understanding any nudge intervention as facilitating learning makes a case for less paternalism over time rather than for more. Accordingly, this focus counters reservations against paternalism

[15] While there may be different views as to whether choice architects should play a role in individual preference shaping, there should not be any disagreement about the importance of also considering effects of nudges on preference learning in the discussion about the acceptability of libertarian paternalism.

[16] 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.

based on slippery-slope arguments (Rizzo and Whitman, 2009b). If behavioral paternalistic interventions aim at increasing autonomy, future interventions will become less likely as individuals will have learned from the previous interventions (or at least future interventions are limited to situations where individuals have problems learning from mistakes).

4. Illustration: "Optimal sin nudges"

This section presents a simple formal model in order to illustrate some potential effects of autonomy-influencing nudges on individual well-being over time. The model is a variant of O'Donoghue and Rabin's (2006) model in which they analyze optimal taxes for "sin goods" assuming that individuals have present-biased preferences. We additionally assume that individuals' perceptions of costs and benefits of the consumption of sin goods can be distorted. We investigate a situation where present-biased preferences but no distorted perceptions are present and lead to overconsumption of sin goods. We assume that choice architects aim at reducing the consumption of these sin goods. Nudges can either reduce present-biased preferences or operate through distorted perceptions, where one case enhances autonomy, while the other decreases it. We sketch possible effects of these nudges on individual subjective well-being, acknowledging that nudges 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. Both types of goods generate "happiness" (in our subjective well-being view) with a decreasing marginal rate of enjoyment. Consuming sin goods has additional future negative consequences. Consumers are identical and their number is normalized to unity. Happiness at time t is given by $h_t = v(x_t; \rho) - c(x_{t-1}; \gamma) + z_t$, where x_t and z_t denote the individual's consumption of sin goods and normal goods, respectively. 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.[17] 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 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 happiness equation can be written as

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

where δ depicts the conventional discount function. In the following, we will assume δ to be unity. In every period, individuals could maximize their happiness given by equation 5 by allocating their income I to the two goods x and z in an optimal way. The prices for both goods are normalized to unity. The per-period income of I is assumed to be large compared to the consumption of unhealthy goods, and individuals do not save or borrow. The optimal allocation of income (x**, z**) maximizes happiness 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 a behavior that maximizes happiness. Following Kahneman (1999), we describe these deviations by assuming that individuals actually maximize their preferences ("decision utility" or "wanting"), although maximizing their actual happiness ("experienced utility" or "liking") would be optimal.[18] In this model, two sources of errors can lead to dissociations between preferences (decision utility/wanting) and happiness (experienced utility/liking). While ε depicts a generic deviation from rationality in equation 2, here we specify these deviations more precisely. The first source of error we consider is the distortion in the perceptions of how beneficial and how costly consumption of the unhealthy good is. Individual perception is described by α. When α = 1, there is no distortion. However, when α

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

[18] Actual happiness is synonymous to Kahneman's (1999) "experienced utility" and Berridge and Aldrige's (2008) "liking". We use actual happiness as benchmark for individuals' interests/well-being. Individual preferences are synonymous to Kahneman's (1999) "decision utility" and Berridge and Aldridge's (2008) "wanting". Preferences, but not actual happiness, can be subject to faulty affective forecasts or other biases (see more extensively Witt and Binder, 2013).

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.[19] Integrating the possibility of distorted perceptions and assuming that δ equals unity, preferences $\hat{u}$ at time t can be described by $\hat{u}(x, z) = v(x; \alpha_\rho\rho) – c(x; \alpha_\gamma\gamma) + z$.

The second dissociation between preferences and happiness is the result of 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 $U(u_t,…u_T) = u_t+\beta\sum^{T}_{\tau=t+1} \delta^{\tau-t}u_\tau$, where u_τ is the individual's decision utility in τ and δ is a conventional discount factor which we assume to be unity. The parameter β depicts a time-inconsistent preference for immediate gratification that reflects impulsivity. While $\beta < 1$ in terms of preferences or wanting, in terms of happiness or liking $\beta = 1$ (see Lades, 2012 for a micro-foundation of β based on the dissociation 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, the preferences 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

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

In every period the individual chooses a consumption bundle (x, z) to maximize equation 6 subject to the budget constraint $x + z \leq I$. Again, prices are normalized to unity, per-period income I is large compared to the consumption of x, and individuals do not save or borrow. When individuals maximize u(x, z) subject to the budget constraint, they allocate their income in a way that maximizes preferences 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 distortions in perceptions lead to deviations from optimal consumption. Note, however, that it is not a priori clear whether distortions in perceptions lead to more or less consumption of the unhealthy good. Perceived costs and

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

benefits of unhealthy consumption can be either higher or lower than actual costs and benefits. Note as well 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 preferences and happiness arising from present-biased preferences. We use the model to illustrate some effects of behavioral policy interventions on individual well-being. Assume that individuals have self-control problems ($\beta < 1$), but no distorted perceptions ($\alpha_\rho = 1$, and $\alpha_\gamma = 1$). Choice architects realize that individuals overconsume the unhealthy good. They decide to nudge individuals towards the choices that make the individuals presumably (see Section 3.1.) better off, i.e. towards consuming x** instead of x*. In equation 6, there are three parameters that nudges can influence to achieve behavioral change: Nudges 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 nudges can influence α_ρ, α_γ, and β for longer than only the current period.

First, assume that to nudge the individuals the policy-maker reduces α_ρ to α_ρ^n ($\alpha_\rho^n < \alpha_\rho$), i.e. the policy-maker reduces the perceived benefits of the unhealthy good (for example by using framing effects and putting junk food in less favorable places in a school cafeteria). Assume that the nudge 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.[20] From the perspective of autonomy-

[20]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.

enhancing paternalism, however, this is an example of manipulation rather than helping the individual making a well-informed, critically reflected decision. The nudge did not help the individual to reduce his well-being reducing error, i.e. $\beta < 1$ is still present. The present-bias is masked or hidden from the view of the individual by another decision making anomaly $\alpha_\rho^n < 1$. What the nudge did is to influence/manipulate choice without influencing the source of the error. If the nudge hides the error, the individual will be less likely to realize similar erroneous behavior in future situations and the individual's capability for making autonomous decisions is decreased.

The problematic character of the described situation becomes particularly obvious when, for any reasons, the decision context changes over time (the individual might, for example, change the school and go to another cafeteria). In the first decision context, the individual did not have the opportunity to understand the actual reasons for the overconsumption of the unhealthy good (actually there was no overconsumption). The individual's present-bias is not reduced so that β is still below unity. Even more problematic, as a result of the nudge at time t, the individual 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 nudge may remove such opportunities. The nudge in period t may thus even increase the individual's 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 nudges are present anymore that could utilize the individual's 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 nudge in the first place ($x' > x^*$).[21] This example illustrates that when acknowledging all dynamic effects, nudges can be detrimental for subjective well-being. This is particularly true when they try to change the outcomes of decision making without influencing the reasons for decision-making errors. This type of nudging is similar to curing the symptoms of a sickness without looking at the underlying reasons of the sickness.

[21] Whether nudges can increase decision-making anomalies over time, however, is essentially an empirical question worth pursuing in future research.

Alternatively, assume that instead of reducing α_ρ to α_ρ^n, the policy-maker 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 temptation (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., 1996), cognitive load (Shiv and Fedorikhin, 1999), visceral states such as hunger and thirst (Lowenstein, 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, 2013). Creating one's own decision environment is an act of autonomous decision making in our definition of autonomy since the possibility to let System 2 decide does not preclude that System 2 defers decisions to the more efficient System 1. 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 thinking rather than lower level thinking (Trope and Liberman, 2003), and make people accountable for their decisions.

Not all of these ways to reduce present-bias are nudges. While nudges change the context in which decisions are made, some of the mentioned interventions aim at changing the individual by training and generating awareness and understanding (Soll et al., 2013). Larrick (2004) argues 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. We like to add the time dimension and argue that learned mental strategies to deal with biases can also benefit future decision-making in the same decision-making context at a later point in time. Such interventions provide possibilities to learn about the underlying reasons of decision-making anomalies and thereby facilitate autonomy and increase β in a more lasting way.

A promising way of autonomy-enhancing interventions is to combine nudges with training exercises, where one, for example, defaults individuals into participating in training sessions that reduce biases.

5. Outlook

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 subjective 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 three major shortcomings: First, it is not clear how choice architects can detect the choices that individuals judge to be in their best interests. Second, libertarian paternalism neglects the fact that behavioral interventions may reduce the individual capability to make critically reflected, autonomous decisions. Finally, libertarian paternalism neglects inter-temporal dynamics that paternalistic interventions can entail.

To remedy these shortcomings the paper has sketched a notion of "autonomy-enhancing paternalism". First, we have argued that it is better to base policy interventions on insights about which choices actually make individuals better off, than on individuals' judgments, which can be mistaken. We suggest that findings from subjective well-being research may provide a clearer notion of what is in the individuals' best interests. The main advantage of using subjective well-being for evaluating libertarian paternalistic policy interventions is that it explicitly deals with the content of what makes individual better off. It is not limited to investigating the way in which individuals make decisions. Secondly, we have explicitly introduced autonomy as an additional constraint that has to be fulfilled so that a nudge can be classified as acceptable. In line with Dworkin (1988), we define autonomy as the possibility to make a critically reflected decision (in System 2). We take seriously the idea of sovereign citizens who want to make well-informed and critically reflected choices. Behaviorally informed paternalism can help them with that, but it should not come at the cost of manipulation. While libertarian paternalism seems to approve of interven-

tions that induce individuals to do the right thing for the wrong reasons, autonomy-enhancing paternalism insists on interventions that help individuals doing the right thing for the right reasons. The importance of autonomy, we have argued, is especially high when analyzing the effects of behavioral interventions in a dynamic framework, where policy interventions can shape preferences and the strength of decision anomalies over time. These various dynamic effects can lead to a reduction of autonomy and, apart from that as well as thereby, reduce individual subjective well-being over time. Without the ability to make autonomous decisions, nudges can put individuals at danger to "learn" being helpless (Binder, 2013) and to lose their ability of pursuing their own happiness (Schubert, 2012). Hence, these aspects have to be considered when engaging in cost-benefit analyses of nudges. In our view, the benefits of behaviorally informed paternalism lies in its intention of helping individuals to overcome their biases and decision-making fallibilities over time and helping them in making 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, and of the 2013 Evolutionary Economics seminar in Jena 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

Amir, O. and Lobel, O. (2008). Stumble, predict, nudge: How behavioral economics informs law and policy. *Columbia Law Review*, 108(8):2098–2137.

Anand, P. and Gray, A. (2009). Obesity as market failure: Could a 'deliberative economy' overcome the problems of paternalism? *Kyklos*, 62(2):182–190.

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

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.

Becchetti, L., Pelloni, A., and Rossetti, F. (2008). Relational goods, sociability, and happiness. *Kyklos*, 61(3):343–363.

Berg, N., Biele, G., and Gigerenzer, G. (2011a). Does consistency predict accuracy of beliefs?: Economists surveyed about psa. Mimeo.

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

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

Beshears J, Choi JJ, Laibson D, Madrian BC.(2008). How Are Preferences Revealed*?. Journal of Public Economics*. 2008; 92(8-9):1787-1794.

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

Binder, M. (2013). Innovativeness and subjective well-being. *Social Indicators Research*, 111(2):561–578.

Binder, M. (2013). Should evolutionary economists embrace Libertarian Paternalism? *Journal of Evolutionary Economics*, in press.

Binder, M., & Coad, A. (2010). An Examination of the Dynamics of Well-Being and Life Events using Vector Autoregressions. *Journal of Economic Behavior & Organization*, 76(2), 352-371.

Binder, M., & Coad, A. (2011). From Average Joe's happiness to Miserable Jane and Cheerful John: using quantile regressions to analyze the full subjective well-being distribution. *Journal of Economic Behavior & Organization*, 79(3), 275-290.

Boyce, C., & Wood, A. (2011). Personality and the marginal utility of income: Personality interacts with increases in household income to determine life satisfaction. *Journal of Economic Behavior & Organization*, 78(1-2), 183-191.

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

Buchanan, J. M. (1991). The foundations for normative individualism. In *The Economics and the Ethics of Constitutional Order*, chapter 18, pages 221–229. The University of Michigan Press, Ann Arbor.

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

Camerer, C. F. (2006). Wanting, liking, and learning: Neuroscience and paternalism. *University of Chicago Law Revue*, 73:87–110.

Camerer, C., Issacharoff, S., Loewenstein, G. F., 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.

Carlin, B. I., Gervais, S., and Manso, G. (2010). Libertarian paternalism, information sharing, and financial decision-making. Mimeo.

Choi, J. J., Laibson, D., Madrian, B. C., and Metrick, A. (2004). For better or for worse: Default effects and 401(k) savings behavior. In Wise, D. A., editor, *Perspective on the Economics of Aging*, chapter 2, pages 81–125. University of Chicago Press.

Christman, J. (2011). Autonomy in Moral and Political Philosophy, *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2011/entries/autonomy-moral/>.

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

de Haan, T., & Linde, J. (2011). Nudge lullaby. Working Paper.

Desai, A. C. (2011). Libertarian paternalism, externalities, and the "spirit of liberty": How Thaler and Sunstein are nudging us toward an "overlapping consensus". *Law & Social Inquiry*, 36(1):263–295.

Diener, E. and Seligman, M. E. P. (2004). Beyond money - toward an economy of well-being. *Psychological Science in the Public Interest*, 5(1):1–31.

Dolan, P., Peasgood, T., and White, M. (2008). Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. *Journal of Economic Psychology*, 29:94–122.

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

Easterlin, R. A. (2003). *Explaining happiness*. Proceedings of the National Academy of Sciences, 100(19):11176–11183.

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.

Frey, B., & Stutzer, A. (2010). Happiness and public choice. *Public Choice*, 144(3), 557–573.

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

Gilbert, D. T. (2007). *Stumbling on happiness*. Harper Perennial, London.

Glaeser, E. L. (2005). Paternalism and psychology. NBER Working Paper No. 11789.

Graham, C. (2009). *Happiness around the world*. Oxford University Press, Oxford/New York.

Graham, C. (2011). *The Pursuit of Happiness*. The Brookings Institution Press, Washington, D.C.

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

Hergenhahn, B. R. and 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-control and the automatic evaluation of tempting stimuli. *European Journal of Social Psychology*, *40*(1), 17-25.

Iyengar, S. S., Jiang, W., and 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. (1999). Objective happiness. In Kahneman, D., Diener, E., and Schwarz, N., editors, *Well-Being: The Foundations of Hedonic Psychology*, pages 3–27. Russell Sage Foundation, New York.

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.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics*, *112*(2), 375-406.

Korobkin, R. (2011). What Comes after Victory for Behavioral Law and Economics. *Unversity of Illinois Law Review*, 2011(5):1653–1674.

Kőszegi, B., & Rabin, M. (2008). Choices, situations, and happiness. *Journal of Public Economics*, 92(8), 1821-1832.

Krueger, A. B. and Schkade, D. (2008). The reliability of subjective well-being measures. *Journal of Public Economics*, 92:1833–1845.

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

Lades, L. K. (2013). Impulsive consumption and reflexive thought: Nudging ethical consumer behavior. *Journal of Economic Psychology*, in press.

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.

Layard, R. (2005). *Happiness - Lessons From a New Science*. Allen Lane, London.

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

Loewenstein, G. and 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.

Madrian, B. C. and Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *The Quarterly Journal of Economics*, 116(4):1149–1187.

Mills, C. (2013). Why Nudges Matter: A Reply to Goodwin. *Politics*, 33(1), 28-36.

Mitchell, G. (2005). Libertarian paternalism is an oxymoron. *Northwestern University Law Review*, 99(3):1245–1277.

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. and Rabin, M. (2003). Studying optimal paternalism, illustrated by a model of sin taxes. *The American Economic Review*, 93(2):186–191.

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

Qizilbash, M. (2009). Well-being, preference formation and the danger of paternalism. Mimeo.

Qizilbash, M. (2011). Sugden's critique of Sen's capability approach and the dangers of libertarian paternalism. *International Review of Economics*, 58(1):21–42.

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. and Whitman, D. G. (2009a). The knowledge problem of the new paternaism. *Brigham Young University Law Review*, pages 103–161.

Rizzo, M. J. and Whitman, D. G. (2009b). 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.

Rosati, C. S. (1995). Persons, perspectives, and full information accounts of the good. *Ethics*, 105(2):296–325.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, 5(17), 61-71.

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

Schubert, C. and 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., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., and 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.

Sobel, D. (1994). Full information accounts of well-being. *Ethics*, 104(4):784–810.

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

Sugden, R. (2008). Why incoherent preferences do not justify paternalism. *Constitutional Political Economy*, 19:226–248.

Sugden, R. (2011). The behavioural economist and the social planner: to whom should behavioural welfare economics be addressed? Papers on Economics & Evolution #1121.

Sumner, L. W. (1996). *Welfare, Happiness, and Ethics*. Oxford University Press, Oxford.

Sunstein, C. R. (2012). The storrs lectures: Behavioral economics and paternalism. Forthcoming in Yale Law Journal.

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

Thaler, R. H. and 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. and 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.

Welsch, H. (2002). Preferences over prosperity and pollution: Environmental valuation based on happiness studies. *Kyklos*, 16(11):1227–1244.

White, M. D. (2008). Behavioral law and economics: The assault on consent, will, and dignity. NEW ESSAYS ON PHILOSOPHY, POLITICS AND ECONOMICS: INTEGRATION AND COMMON RESEARCH PROJECTS, Gerald Gaus, Christi Favor, Julian Lamont, eds., Stanford University Press, Forthcoming. Available at SSRN: http://ssrn.com/abstract=1274444.

Wilson, T. D. and Gilbert, D. T. (2005). Affective forecasting - knowing what to want. *Current Directions in Psychological Science*, 14(3):131–134.

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

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