ERASMUS SCHOOL OF PHILOSOPHY

Research Master Thesis

Behavioral Paternalism: What is it? Can it be Defended?

By: Lior Nissim Grinman

Supervisor: Prof. Dr. Jack Vromen

Advisor: Dr. Conrad Heilmann

Word count: 23,230

Number of EC: 30

Date of completion: July 6, 2020

Table of Contents

Introduction

Economists have been somewhat suspicious of objective conceptions of the good, warning these can lead to paternalism.[1] Indeed, some of the best advocates of so-called 'Neo-Classical' economics often justify their advocation of free markets using the language of subjective judgment of the good.[2] In a like manner, welfare economists often try to avoid paternalistic actions by resorting to people's judgments about what is good for them.[3] This is often reflected in using the satisfaction of their preferences as a welfare criterion. Thus, the question that welfare economics is traditionally concerned with is not what people should choose. Rather, it is how to satisfy their preference to one choice over another; and when different people have conflicting choices, then how to balance or aggregate these.

However, in the last few decades a research program called Behavioral Economics has been challenging the idea that people always choose what they prefer.[4] To explain the distorted relationship between preferences and choices, behavioral economics invokes the concept of reasoning failures. Just as markets can have failures that prevent them from arriving at efficient allocations, so people can have failures in their reasoning that prevent them from choosing as they truly prefer. In reply to this challenge, a new area of welfare economics emerged that purports to reconcile the findings of behavioral economics with welfare economics. This new area is now called by some Behavioral Welfare Economics. Although behavioral welfare economics is a collection of many different views, one approach particularly grabbed much attention in recent discussions. This approach, as I shall call it in this thesis, is *Behavioral Paternalism*. Having accepted the claim that people do not always choose as they prefer, behavioral

[1] Robert Sugden, *The Community of Advantage: A Behavioural Economist's Defence of the Market*, 2018, 6, 12; Johanna Thoma, "Merely Means Paternalist ? Prospect Theory and ' Debiased ' Welfare Analysis," 2019, 1.

[2] Partha Dasgupta, "Positive Freedom, Markets and the Welfare State," *Oxford Review of Economic Policy* 2, no. 2 (1986): 25–36.

[3] Historically speaking, there is a difference between welfare economics in the neo-classical period and the one which came after it – what is now known as 'New Welfare Economics'. My discussion of welfare economics resembles more of the latter. However, as this is not a historical project, I refrain here from going further into discussing these two historical projects and refer interested readers to Blaug & Backhouse (1986).

[4] To be sure, I am not suggesting that all behavioral economics is concerned with is disproving the relationship between preferences and choices. But for the purposes of this discussion, this part of behavioral economics is most relevant.

paternalists want to help people to choose better. That is, in the same way they would choose had they not been subjected to reasoning failures. But by intervening with people's choices, behavioral paternalists introduced a new justification for paternalism in economics. Just as markets sometimes need interventions to be efficient, so people sometimes need to be interfered with their choices to satisfy what they truly prefer. Thus, behavioral paternalism presumably salvages the appreciation for subjective judgment of the good that was salient in neo-classical economics but with a paternalistic aspect.

Yet, how can behavioral paternalists interfere with people's choices to make them choose one thing over another, while still holding on to the claim that by doing so people will choose the alternative that made them better off, according to their own subjective standard? Or put differently, how do we make someone better off, "as judged by himself" (to quote some of the leading behavioral paternalists)[5], by considering only certain preferences as relevant for the welfare domain? This thesis is an attempt to answer these questions. To be sure, this inquiry is about evaluating the ability of behavioral paternalistic policies to increase each person's subjective well-being, when considered in isolation. Thus, it leaves out of the analysis questions about social well-being. It also leaves out cases where behavioral policies are used to achieve some public policy objective and not improve well-being, such as the reduction of air pollution or tax evasion.

The thesis is organized as follows. Chapter 1 shall provide the context for the discussion. It discusses what is standard welfare economics, the role rationality plays in it, how standard welfare economics is challenged by behavioral economics and how different approaches in behavioral welfare economics respond to this challenge. I claim that we can think about the challenge as twofold. First, some of the experiments and observations in behavioral economics impugn the assumption that people behave in a rational way. Thus, it seems that people do not always choose what they prefer. I call this the descriptive challenge. But if choices do not always reflect people's preferences, then the implication of this is that using preference-satisfaction as a welfare criterion

[5] Richard H. Thaler and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, 2008.

becomes problematic as well. I call this the normative challenge. These challenges are then used to distinguish behavioral paternalism from other approaches. I claim that we can think about behavioral paternalism as a behavioral welfare program that accepts the descriptive challenge, while at the same time rejects the normative one. Following these preliminaries, the chapter identifies two salient characteristics of behavioral paternalism. First, its use of the satisfaction of purified preferences as a welfare criterion. Purified preferences are the preferences people would reveal had they been fully rational and not subject to reasoning failures. Reasoning failures are explanations for the violations of rationality. I name four such failures: limited technical ability, limited imagination, limited willpower and limited objectivity. Furthermore, behavioral paternalism purports to satisfy purified preferences to increase people's well-being. Therefore, the satisfaction of purified preferences also becomes the constituent of well-being in this approach. The second salient characteristic of behavioral paternalism has to do with its interference with choices. Because it interferes with choices, behavioral paternalism has a clear paternalistic aspect.

But how does behavioral paternalism different than traditional paternalistic accounts? Or is it the same thing? Chapter 2 shall deal with these questions as well as issues of definitions of paternalism. It reviews prominent general definitions for the term as well as definitions for paternalism in economics. I argue that these definitions are unsatisfactory in describing behavioral paternalism. I then propose a new definition of behavioral paternalism. In brief, I conclude that a policy is behavioral paternalistic if, a) it interferes with choices; b) justified by reasoning failures; c) done for the individual's own good; and d) while not infringing upon his liberty, narrowly understood. I then examine to which cases this definition can be extended. It turns out that there are three cases in which behavioral paternalism purely scores. Those are: means-paternalism, broad-paternalism and welfare-paternalism.

Chapter 3 uses the previous arguments to answer the central question: how can we interfere with someone's choices to make him better off, from his own subjective perspective? I claim that because behavioral paternalism purports to satisfy only purified preferences and also to be means-paternalist, then it equates purified preferences with people's own goals. Thus, to evaluate whether behavioral paternalists can interfere with people's choices to make them better off by their own standard, we

need to examine whether it is possible to establish the claim that the satisfaction of purified preferences can really define well-being; that satisfying purified preferences is equal to helping people achieve their own goals. In other words, to answer the main question we need to justify the use of purified preferences. However, I identify two main challenges in the literature to achieve this. First, it is claimed that behavioral paternalists think (even if only implicitly) about purified preferences as something that really exists in people's minds in some psychological sense. However, some of our daily experiences and psychological observations should make us skeptical about this idea. Second, even if we assume that purified preferences exist, to use them in a way that would truly improve people's overall well-being seems untenable. The reason is that this requires knowledge about people's preferences, reasoning failures and the context in which they appear. But this seems to require too much. I then suggest a reply to these challenges. First, I claim that behavioral paternalists need not think about purified preferences as something tangible that exists in people's minds. Rather, they can (and even seem to) think about purified preferences as counterfactual-ifs. Purified preferences can be thought of as a mere useful technical tool that helps to explain what preferences would be observed had people been fully rational and not subject to reasoning failures. Second, I argue that behavioral paternalists do not need to counter every possible trade-off in the agent's well-being. They only need to be able to do better than the agent, and at least most of the time. More specifically, I suggest four reasons why it seems plausible that they can achieve this. I also argue that in cases where we suspect that behavioral paternalists would not do better, we can use different safeguards to ensure that the agent's overall well-being will not be decreased.

Chapter 1: Normative and behavioral economics

This chapter asks what Behavioral Paternalism is. It delineates behavioral paternalism by distinguishing it from other approaches in behavioral welfare economics. In doing so, it also develops a typology of the challenges that welfare economics aims to solve in response to some of the observations of behavioral economics.

The chapter is organized as follows. First, it describes Standard Welfare Economics (SWE), its welfare criterion, and the role rationality plays in it. In brief, SWE's welfare criterion uses the notion of satisfying individual preferences. But since preferences are assumed to be reflected in choices, they should take into account everything that people care about. As such, the axioms of rational choice become justified as properties of sound reasoning about choice. This is explained in section 1. Second, the chapter reviews the challenge to SWE's welfare criterion put by Behavioral Economics. The challenge is twofold. First, behavioral economics' empirical observations challenge the assumption that people display behavior that complies with the axioms of rationality. I refer to this as the descriptive challenge. Second, if choices do not always reflect people's preferences, then the implication of this is that using preference-satisfaction as a welfare criterion becomes problematic as well. I refer to this as the normative challenge. This is done in Section 2. Lastly, the chapter examines how different approaches of welfare economics purport to deal with these challenges, focusing mainly on behavioral paternalism. In line with the typology I develop, I characterize behavioral paternalism as a behavioral welfare economics approach that takes the first challenge seriously, but that rejects the second challenge. This is done in Section 3.

1. Standard welfare economics and rationality

Preferences are a central concept in modern economics. In positive economics, they are used both to explain and predict behavior. In normative economics, preferences are used to measure welfare. More specifically, economists rely on the satisfaction of individual preferences as a welfare criterion.[6]

[6] Robert Sugden, "On Nudging: A Review of Nudge: Improving Decisions about Health, Wealth and Happiness by Richard H. Thaler and Cass R. Sunstein," *International Journal of the Economics of Business* 16, no. 3 (2009): 366; Robert Sugden, "The Community of Advantage: A Behavioural Economist's Defence of the Market," *The Community of Advantage: A Behavioural Economist's Defence of the Market*, no. November 2019 (2018): 6.

SWE can be very much characterized as a project that refrains from deciding between different conceptions of the good when doing welfare analysis. Indeed, welfare economists usually desire to allow people to use their own subjective judgments for what is good for them, thus, eschewing away paternalism as much as possible.[7] However, SWE does not only try to avoid references to an objective conception of the good. It also wants to refrain from anything that is 'internal' or unobservable of the agent, such as mental states.[8] Thus, we can define SWE as a welfare approach that i) uses a subjective judgment of the good; while ii) not require the analyst to make internal observations.

Yet, one can notice that preferences are, by definition, a mental state. Therefore, how can one use them to evaluate welfare, while at the same time claim to refrain from mental states? To achieve these two goals, SWE makes a further assumption – that preferences are reflected in choices. This is often known as the Revealed Preference Method.[9] Put more formally, if alternatives A and B are equally available to the agent and he chose A over B, then we can assume he prefers A over B as well. Crucially, a welfare policy that gives the agent alternative A over B shall be welfare improving for him.[10]

We can note that since a preference is a subjective judgment, by definition, using the satisfaction of preferences satisfies the requirement for a subjective judgment of the good. Similarly, assuming a relationship between preferences and choices allows the welfare analyst to infer from choice data regarding the welfare of the agent without referring to any internal observations. Thus, the two requirements of SWE mentioned above are satisfied.

But what are preferences? There are various interpretations of this question. However, in economics, perhaps the most common ways to think about preferences are that they

[7] I will not elaborate on paternalism in this chapter. A discussion regarding the term shall take place on chapter 2.
[8] Daniel M. Haybron and Anna Alexandrova, "Paternalism in Economics," *Paternalism: Theory and Practice*, 2011, 157–77; Faruk Gul and Wolfgang Pesendorfer, "The Case for Mindless Economics," in *The Foundations of Positive and Normative Economics: A Handbook* (Oxford University Press., 2008), 40.
[9] Haybron and Alexandrova, "Paternalism in Economics".
[10] Gul and Pesendorfer, "The Case for Mindless Economics," 24.

are either dispositions to choose one alternative over the other, or that they are the choices that reveal them and nothing more.[11] Either way, preferences are reflected in choices. Yet, an important implication of assuming a link between preferences and choices is that if preferences truly reflect choices, then they should take into account everything the agent judges to be relevant for the decision. In Daniel Hausman's (2012) words, preferences should be thought of as 'Total Comparative Evaluations'.[12] Otherwise, if preferences do not take into account everything that is relevant to the agent, then it will be problematic to use them to explain, predict and design welfare policies. To see why, let us assume that the concept of preferences does not take into account everything the agent judges to be relevant. Now suppose that, say, moral commitments are excluded from the term 'preferences'. Then, first, it will be hard to explain choices using preferences as it might be the case that a moral commitment dictated the choice and not the preference. Second, if some moral commitment conflicts with a preference, then adhering to that moral commitment will presumably make the individual worse-off since preference-satisfaction is used as a welfare criterion. However, this seems to be counter-intuitive.[13]

Importantly, the connection between preferences and choices also makes economics a theory of choice. As with any account of choice, economics should attempt to explain and predict people's choices by their reasons. Therefore, as an account of choice, economics need to show how the factors it takes to be reasons can justify choices. To achieve this, economists postulate several axioms governing preferences. Consequently, these axioms can be read as imposing conditions on a rational choice. Put differently, one way to interpret economists' use of such axioms is by asking, what restrictions should be imposed on choices by any consistent set of reasons? In the case of the theory of choice in economics, the answer to this question is found in complying with the axioms imposed on preferences. Preferences that comply with these axioms are often

[11] Richard Bradley, *Decision Theory with a Human Face* (Cambridge University Press, 2017), 45.
[12] Daniel M. Hausman, *Preference, Value, Choice, and Welfare* (Cambridge University Press, 2012). Hausman further assumes that only preferences, together with beliefs, dictate choices. My use of his account here is just to elucidate the idea that if preferences are to be reflected in choices in some way, then they should take into account everything the agent judges to be relevant for the decision. Therefore, I do not necessarily embrace his entire account of preferences and beliefs.
[13] Hausman (2012). I really heavily on Hausman in explaining this part and the following one.

regarded as 'rational'. Or put differently, choices that reflect such rational preferences can be seen as motivated by reasons.

Although there is disagreement in economics regarding what axioms are needed for rationality, most economists agree on the following assumptions:[14]

i) People have well-defined preferences and they make decisions that maximize these preferences
ii) These preferences accurately reflect the true costs and benefits of the available alternatives
iii) In cases of uncertainty, the agent has well-formed beliefs regarding how the uncertainty will resolve

Let us explain these assumptions. First, the assumption about well-defined preferences usually consists of several axioms. The most common ones are the following. [15] Firstly, there are two axioms regarding internal consistency: *Completeness* and *Transitivity*.[16] Completeness means that facing two alternatives A and B, an agent will either strictly prefer A to B, or B to A, or otherwise be indifferent between them. Completeness also uses a universality requirement – that is, that *any* two alternatives in the universe of possible alternatives are ordered in this way. Transitivity entails that for any three alternatives A, B and C, if the agent prefers A to B and B to C, then he must prefer A to C as well. These two axioms set out to prevent internal contradictions and ensure that preferences are consistent.[17]

Although they are usually not explicitly referred to as axioms, economists often rely on two additional axioms: *Stability* and *Context-independency*. Stability stipulates that preferences do not change sharply in the short-run or in short periods of time and do not vary randomly, at least not significantly. Context-independency assumes that

[14] Colin Camerer et al., "Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism,'" in *University of Pennsylvania Law Review*, 2003, 1214–15.
[15]In defining these axioms I rely on Hausman, *Preference, Value, Choice, and Welfare*, pt. 1; Sugden, *The Community of Advantage: A Behavioural Economist's Defence of the Market*, 5.
[16] I am loosely following Whitman & Mario (2015, p. 412), in defining internal consistency here.
[17] A third axiom if often used as well, which is *Independence of Irrelevant Alternatives*. The idea is that a preference relation between A and B should not be changed or reversed if a third option C is introduced. Whitman & Rizzio (Ibid.) claim that the independence of irrelevant alternatives axiom can be implied by Completeness. In any case, I will not discuss this axiom anymore in the following.

preferences are not affected by irrelevant factors of the decision problem. Or put more formally, that whether an agent prefers A to B remains stable across contexts. This axiom connects preferences in one set of alternatives to another set of alternatives. Thus, it ensures consistency across different contexts.

Second, there is the assumption about weighing gains and losses accurately. To illustrate with a text-book example, suppose that Bob is asked to make a monetary bet based on the color of a ball drawn from an urn. Bob knows the distribution of colors (i.e., how many blue, red and yellow balls there are in the urn). We can say that Bob complies with the second assumption if he has calculated correctly the possible monetary gains and losses based on the prevalence of the balls in the urn.

Third, there is the assumption about well-formed beliefs in the face of uncertainty. One example of a well-formed belief is that as new information comes in, the agent updates his beliefs according to Bayes' law (that is, update probabilistic assessments according to the new information). To illustrate, suppose now that Bob is asked to make another bet. In that new bet, he wins if he draws a blue ball, having drawn a yellow one already. If Bob is successful in updating his posterior probabilities (i.e., the new probabilities of drawing balls from the urn) based on the prior ones, then he has a well-formed belief.

Henceforth, when referring to 'rationality' or someone being (ir)rational, I shall use the term with regards to these axioms.

In other words, if preferences are Total Comparative Evaluations, then the axioms above become justified as properties of sound reasoning about choice; as something that would be reflected in the behavior of a rational person. Or put differently:

> "if a rational choice is one that is justified by sound and relevant reasons, and if reasons are to take the form of total comparative evaluations of the feasible options, rational choice is possible in general only if those evaluations are complete, transitive and context independent. In this sense, Hausman's interpretation of 'preference' offers an

explanation of why the axioms of choice theory are treated as principles of rationality."[18]

Nonetheless, the idea of using preference-satisfaction as a welfare criterion stems from various motivations. That is, the answer to the question '*Why preference-satisfaction matters to welfare?*' has had different interpretations in history. Chief among these are *Happiness, Freedom* and *Well-Being*.[19] The happiness interpretation assumes that happiness itself – or hedonic experience as it is often called – is welfare improving. That is, that more happiness shall lead to more welfare. Even though that on this approach satisfying a preference is not equal to happiness, it is assumed that satisfying a preference could be the best way to make people happy. [20] Thus, revealed preference is assumed to be a reliable index for happiness.[21] The freedom approach emphasizes freedom of choice as welfare improving.[22] Thus, on this interpretation, satisfying preferences is only valuable because of the connection between preferences and choices and because it is good to reinforce freedom of choice. Lastly, the well-being approach emphasizes well-being. Furthermore, it defines well-being as the satisfaction of preferences. Thus, satisfying preferences increase one's well-being.[23]

Nonetheless, the advantage of using preference-satisfaction as a welfare criterion is that economists could remain silent on these interpretations. This is because as long as preferences can be assumed to comply with rationality, to be reliably reflected in choices, each approach could still link preferences to what mattered to it most (i.e., freedom, happiness or well-being). Thus, economists could use preference-satisfaction

[18] Gerardo Infante, Guilhem Lecouteux, and Robert Sugden, "Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioural Welfare Economics," *Journal of Economic Methodology* 23, no. 1 (2016): 11.
[19] This part largely follows the account of McQuillin & Sugden (2012, p. 555).
[20] Hausman, *Preference, Value, Choice, and Welfare*, 80.
[21] To be clear, the happiness approach acceptance of preference-satisfaction was before the era of behavioral economics. This clarification is important because the happiness approach as it is known today is often seen as an alternative to the preference-satisfaction criterion.
[22] Also, assuming that it is good that each individual has the freedom to make decisions to the extent it effect only him.
[23] It might be objected that the happiness approach is also a well-being approach (often called Hedonism). To address any confusion, the so-called well-being approach might be identified with a desire-satisfaction account of well-being, which is different from Hedonism. I refer to it as 'the well-being approach' here to be coherent with literature. In Section 3, we shall define it as 'Behavioral Paternalism', and then I shall use that term instead. For more details on these differences in well-being approaches, I refer the reader to Derek Parfit, "What Makes Someone's Life Go Best," in *Reasons and Persons*, 1986, 1–12; Guy Fletcher, *The Philosophy of Well-Being: An Introduction* (Routledge, 2016).

as a welfare criterion without coming to an agreement on these different interpretations.[24]

Before moving on, let us summarize the discussion so far. I have described SWE as a welfare approach that purports to use a subjective judgment of the good. To achieve this, it uses preference-satisfaction as a welfare criterion. But SWE also desires to avoid making internal observations of the agent. To accomplish this, SWE uses the revealed preference method in which preferences are reflected in choices. But because of the connection between preferences and choices, economics becomes a theory of choice. As such, it should justify choices by their reasons. In that context, economists postulate several axioms that govern preferences. Preferences that comply with these axioms are considered as 'rational.' More specifically, the following assumptions are commonly used: completeness, transitivity, stability, context-independency, accurately reflect gains and losses and having well-formed beliefs in cases of uncertainty. Lastly, although there are different interpretations of the question, 'why preference-satisfaction matters', economists can stay silent on these as long as preferences can be assumed to be reflected in choices.

In the last few decades, however, the rationality axioms are challenged by a research program known as Behavioral Economics. In the following section, I elaborate on this challenge.

2. The challenge from behavioral economics

Rationality is much used in economics. In providing explanations and predictions of behavior, it is assumed that people behave (or at least behave 'as-if') in a way that reflects rational preferences. Thus, explanations in economics are regarded as good if they explain phenomena in terms of the rational choices of individuals. Furthermore, in designing welfare policies, economic institutions and policymaking are justified if their outcomes score highly on the preference ordering of individuals. [25]

[24] McQuillin and Sugden, "Reconciling Normative and Behavioural Economics: The Problems to Be Solved," 555.
[25] McQuillin and Sugden, 553.

Yet, in the last few decades there has been a change. A research program called Behavioral Economics (BE) had been producing mounting work that challenges the validity of rationality. One of the key findings of BE is what is often known as 'systematic anomalies'. These are observations that individuals, in certain situations and contexts, systematically deviate or do not behave 'as-if' their preferences are rational. To illustrate, let us review some of the well-known examples for these anomalies:[26]

- *Framing effects*. Framing effects are the observation that describing a choice problem in two different but logically equivalent ways often leads to significant distinct choices. In other words, although the alternatives remain the same, the way of describing or 'framing' the choice problem, affects the outcome. Redelmeieer *et al*. (2012)[27] is a good illustration of this. They describe that people are more inclined to support a medical procedure if they are told there is a 90% of success. Yet, their preference will be reversed if they are told that there is a 10% of dying from that same procedure. Logically, of course, these probabilities are complementaries. That is, the choice in both is the same between 'a procedure with 90% of success, 10% of dying'; or 'not doing the procedure at all'. Framing effects can be considered as a violation of completeness, transitivity and context-independency. The framing effect is sometimes related to the following two anomalies as well: endowment effect and status quo bias.
- *Endowment effect*. This anomaly occurs when the agent's willingness to pay (WTP) for a good differs significantly from his willingness to accept (WTA) money for that same good. Kahneman *et al.* (1991) are illustrative of this.[28] They show that subjects differ significantly in their WTP and WTA, respectively, if they were given a coffee cup and need to sell it, or were in need to buy it instead from the experiment conductors. Logically, the choice is the same between 'money' or 'coffee'. The endowment effect violates completeness.

[26] Strictly speaking, this is not a list of anomalies. A more proper term for this is a list of theories explaining how preferences that violate the axioms of rationality come about. However, sometimes in the literature these theories are used to explain the violations of rationality. Whitman and Rizzo (2015, pp. 413-414) are an example of this. They present a useful review of these violations, and I am largely following them in my exposition here. Also, note that this is a non-exhaustive list, of course.

[27] Donald A Redelmeieer, Paul Rozin, and Daniel Kahneman, "Cognitive and Patients' Decisions," *Jama* 270, no. 1 (2012): 72–76. An additional prominent phenomena related to the framing effect is *loss aversion* – i.e., that people are more inclined to avoid losses than prefer equal gains.

[28] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler, "The Endowment Effect, Loss Aversion, and Status Quo Bias," *Journal of Economic Perspectives* 5, no. 1 (1991): 193–206.

- *Status quo bias*. This anomaly occurs when, facing with two alternatives (or more), people would usually prefer to stick with the default option. An example of the status quo bias is the 'save more tomorrow' program that enrolls workers in a savings program by default.[29] People can opt-out if they want. Yet, the program was proven to increase the average saving rate, nonetheless. Arguably, it might be objected that the choice between opting-in and opting-out of something may not always necessarily be equivalent. Opting-out of something can be, at times, harder. Think of the last time you have joined a service like a mobile provider. Often, the call center picks up the call within a few minutes and connects you to the service quite easily. Disconnecting from the service, however, can be the opposite experience at times – longer waiting times on the line and paying cancelation fees or switching costs. Yet, behavioral economists claim that in cases such as *save more tomorrow*, "the cost of returning-in a form is trivial". [30] In other words, they consider the cost to be small enough such that the choice is the same and therefore should not be altered by changing the default. If this is convincing, then the status quo bias can point to an anomaly as well. The status quo bias can be considered a violation of completeness, transitivity (or both) and context-independency.
- *Hot-cold states.* 'Hot states' are emotional situations (e.g., fear, excitement) in which the agent is more inclined to make certain choices, such as purchasing certain goods, more than he would have in a 'cold state' (e.g., being calm and sober).[31] This anomaly can be considered as a violation of completeness and, or, transitivity.
- *Hyperbolic time discounting*. Simply put, this anomaly is the observation that individuals behave as if they have a declining discount rate function instead of one that remains the same (i.e., exponential discounting).[32] That is, if an agent prefers a larger amount of good (A) at time t+1 to a smaller amount of that good (B) at time t, he might reverse that ranking if t and t+1 are close enough to the present time. It is easy to see

[29] Richard H. Thaler and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, 2008, 105–19.
[30] Cass R. Sunstein and Richard H. Thaler, "Libertarian Paternalism Is Not an Oxymoron," *University of Chicago Law Review*, 2003, 1171.
[31] George Loewenstein, "Emotions in Economic Theory and Economic Behavior," *American Economic Review* 90, no. 2 (1999): 256–60.
[32] George Ainslie, "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control," *Psychological Bulletin* 82, no. 4 (1975).

that hyperbolic discounting could violate completeness or transitivity (or both) and stability as well.[33]

This evidence can pose a challenge to economics. First, it impugns the description of people as behaving rationally. Or put differently, the evidence challenges the assumption that people (always) behave 'as-if' their preferences are rational. As a consequence, it is claimed that people do not always choose as they prefer. Call this *the descriptive challenge*. Second, the violations of rationality also challenge the ability to design welfare policies and evaluate their outcomes successfully. If a welfare policy should be justified by preference ranking, and preferences are assumed to be reflected in choices, then we should assume that the link between choices and preferences is stable. That is, that the agent chose A because it was truly preferred and that by satisfying his preference for A we are improving his welfare. However, the observations of BE imply that people's choices do not necessarily reflect their preferences. Consequently, if people's choices do not always reflect their preferences, then using preference-satisfaction as a welfare criterion becomes problematic as well. I call this *the normative challenge.* The two challenges are not completely unrelated. Since SWE uses preference-satisfaction as a welfare criterion, and preferences are assumed to comply with all axioms of rationality we reviewed in Section 1, then biting the bullet of the descriptive challenge can entail a problem for SWE as well. Put differently, if individual behavior does not always reflect rational preferences such that preferences are not reflected in choices anymore, then a method that tries to infer from choices about preferences to design welfare policies can be problematic as well.

To elucidate these challenges, consider a well-known experiment by Read & Van Leeuwen (1998).[34] In their experiment, participants were asked in advance to decide between healthy or unhealthy snacks that will be delivered later. The results showed that the participants' decisions differed based on the time they were asked to choose. If they were asked to choose late in the afternoon when they were hungry, they were more

[33] Hyperbolic discounting can also be considered as a violation the *Stationarity* assumption – that is, the assumption that preferences do not change over time. But to be sure, the story of time discounting is more complicated than presented here as there are various approaches to explain the anomalies. For simplicity, however, I shall not review these here and refer interested readers to Frederick *et al* (2002).
[34] Daniel Read and Barbara Van Leeuwen, "Predicting Hunger: The Effects of Appetite and Delay on Choice," *Organizational Behavior and Human Decision Processes* 76, no. 2 (1998): 189–205.

inclined to choose the unhealthy snack. Conversely, if they were asked straight after they had lunch, they were more inclined to choose the healthy snack. What the descriptive challenge means here is that the results show an axiom(s) of rationality is violated. In this case, whether the participants choose a healthy or unhealthy snack depended on a context, thus violating the transitivity and context-independency axiom. The normative challenge means that a welfare analyst cannot judge if the decision to eat an unhealthy or healthy snack was really the one preferred by the individual when observing his choices. In other words, while the descriptive challenge means that an anomaly with regards to rationality had occurred, the normative challenge also means that it is not possible to design a welfare policy due to this anomaly. To some readers, this might seem like a redundant distinction at first. However, in the next section we shall see why this terminology can be helpful to distinguish between different responses in welfare economics to the anomalies.

3. What is behavioral paternalism?

How do economists respond to the above challenges? The answer is that it depends. In Section 1, we saw that interpretations to the question *'Why preference-satisfaction matters to welfare*?' can vary. But as long as people were assumed to behave rationally such that preferences could be assumed to be reflected in choices, economists could use preference-satisfaction as a welfare criterion and remain silent on these interpretations. However, the descriptive and normative challenges virtually suggest that this may not be possible anymore. Accepting both of them as true implies that an alternative approach (or at least some revision) is needed for the welfare criterion. Consequently, adherents of the different interpretations purport to find ways to reconcile the descriptive challenge with a normative account. That is, to develop a new welfare criterion that takes into account the observations of BE.

For most of this section, I shall focus on the response of the well-being approach to the descriptive and normative challenges. However, I begin by reviewing the replies of the other approaches. Such discussion is helpful in distinguishing the well-being approach by locating it on a spectrum of responses in welfare economics.

First, it should be noted that not all economists accept the descriptive challenge (although they may not be the majority). Call this the 'skeptic view'. One attempt to

deal with the descriptive challenge is to claim that the anomalies are not robust. For example, that inconsistent behavior decay with the experience individuals gain and if a decision task is repeated enough times, behavior that reflects rational preferences would be observed.[35] Thus, this approach would not need a revision for the normative part. Second, under the freedom interpretation, choices are not valued as a means to reveal preferences, but because of their intrinsic value of being one's choices. Therefore, choices are valued because they have the potential to promote one's freedom of choice. Thus, the freedom interpretation suggests that it is welfare improving for individuals to choose for themselves regardless of whether that choice can be considered rational. However, this approach still needs some refinement for the normative part. This is because it accepts the normative challenge – i.e., that designing welfare policies based on preference orderings becomes problematic. Consequently, some attempts to refine the normative criterion in the freedom approach suggest that the welfare analyst needs to ask to what extent the outcomes bestow individuals with more freedom to choose.[36] Thirdly, there is the happiness approach. Proponents of this approach accept both challenges as well. One prominent way for them to reply to these challenges is to make the welfare criterion maximize happiness by focusing on hedonic qualities. This is often known as the 'Experienced Utility' approach.[37]

Figure 1. shows the different approaches and their response to the respective challenges. To be sure, I shall not discuss the above approaches anymore. I continue by discussing the well-being approach and focus on it in the remainder of this thesis. This approach has grabbed a lot of attention in recent discussions due to its reply to the aforementioned challenges. Perhaps this should not be too surprising. The view of well-being as preference-satisfaction is considered the mainstream in modern welfare economics.[38]

[35] James B Kliebenstein et al., "Resolving Differences in Willingness to Pay and Willingness to Accept," *The American Economic Review* 84, no. 1 (2016): 255–70; Graham Loomes, Chris Starmer, and Robert Sugden, "Do Anomalies Disappear in Repeated Markets," *Economic Journal* 113, no. 486 (2003): 153–66; Charles R. Plott, "Rational Individual Behavouir in Markets and Social Choice Processes: The Discovered Preference Hypothesis," in *Rational Foundations of Economic Behaviour*, 1996. To be sure, some do claim that it seems plausible that the evidence is robust (cf. McQuillin &Sugden, 2015, p. 554).
[36] McQuillin and Sugden, "Reconciling Normative and Behavioural Economics: The Problems to Be Solved"; Sugden, *The Community of Advantage: A Behavioural Economist's Defence of the Market*.
[37] D. Kahneman, P. P. Wakker, and R. Sarin, "Back to Bentham? Explorations of Experienced Utility," *The Quarterly Journal of Economics* 112, no. 2 (1997): 375–406; George Loewenstein and Emily Haisley, "The Economist as Therapist: Methodological Ramifications of 'Light' Paternalism," *The Foundations of Positive and Normative Economics: A Hand Book* 1 (2007): 12–14.
[38] Hausman, *Preference, Value, Choice, and Welfare*, pt. 2.

But, perhaps, it is more because of the response itself that the well-being approach grabs so much attention. As we shall see, this approach advocates paternalism, something that, as mentioned in the introduction, was quite foreign to economics. In any case, the well-being approach deals with the challenges above by accepting the descriptive challenge while at the same time rejecting the normative one. It does so by a twofold claim. First, it claims that the anomalies observed by BE imply that choices do not necessarily reflect individuals' well-being. Second, it takes the path of reconstructing the conception of well-being *that would govern the individual's behavior had he been fully rational*.[39] Let us explain these claims.

Figure 1. responses of the different approaches to the challenges

Challenges/Approaches	Descriptive	Normative
The skeptic view	Not accepting	Not accepting
Happiness	Accepting	Accepting
Freedom	Accepting	Accepting
Well-being	Accepting	Not accepting

The first claim is supported by simply accepting the descriptive challenge. That is, by accepting that we cannot merely infer from choices on one's preferences. However, to support the second claim, the well-being approach accepts the concept of *reasoning failures*. The idea is that the anomalies observed on the descriptive level can be explained by a lack of full rationality from the individual's side and that such violations of rationality are caused by reasoning failures. There exist different classifications of these failures with different names for them.[40] However, a useful classification is Le & Grand & New (2015) [41], which names the following four:

- limited technical ability
- limited imagination
- limited willpower
- limited objectivity

[39] McQuillin and Sugden, "Reconciling Normative and Behavioural Economics: The Problems to Be Solved," 556.
[40] Infante *et al.* (2016) refer to them, for example, as 'reasoning imperfections' and name the following: limitations of attention, information, cognitive ability, or self-control. Thaler & Sunstein (2008), also use a similar list.
[41] Julian Le Grand and Bill New, *Government Paternalism: Nanny State or Helpful Friend?* (Princeton and Oxford: Princeton University Press, 2015), chap. 5. I rely heavily on Le Grand & New in describing these reasoning failures.

Let us elucidate these terms. First, limited technical ability encompasses difficulties with formal, technical analysis and perception. It usually results from presenting information in a complex way. For example, assessing probabilities correctly belongs to this category. Second, limited imagination has to do with people's limited ability to imagine or predict how their preferences would change due to some experience, at times in the far future. For example, suppose that when Carol is twenty years old she needs to decide if she wants to open a savings account for her retirement. Limited imagination can make it difficult for Carol to imagine how difficult her life would look like when she is much older and retire with no money saved. Third, limited willpower has to do with people's weakness of will. This reasoning failure occurs in situations in which people know what they prefer, or perhaps what they prefer in the long-term, yet choose against that preference. For instance, suppose that John wants to lose weight and knows what he needs to eat to achieve that goal. Yet, every time he has the opportunity, John orders a desert because he 'simply cannot help it'. John's choice to eat a desert can be considered as a weakness of will. Lastly, limited objectivity is a failure to behave without biases, that is, objectively. This failure encompasses many biases and I shall not offer an exhaustive list here. But to illustrate, consider the confirmatory bias. Someone has this bias if she misinterprets information to confirm her hypothesis. Suppose that Anne, for many years now, believes that eating gluten causes her pain and thus chooses not to eat gluten products. Anne completely changed her eating lifestyle so that she would not have to eat gluten. However, suppose that now a new medical theory is introduced that does not support Anne's hypothesis regarding the relationship between gluten and her pain. However, Anne finds a way to interpret the new theory to support her belief that gluten causes her pain nonetheless. Thus, she chooses to continue in her gluten-free diet. We can say that Anne has limited objectivity.

Having explained the violations of rationality by reasoning failures, the well-being approach considers choices that are affected by such failures as mistakes. The claim is that choices made because of reasoning failures are not the choices individuals truly wanted to make – i.e., the ones that would satisfy their true preferences and therefore increases their well-being. Consequently, the well-being approach claims that by intervening with people's choices such that they will not make mistakes anymore it is possible to increase their well-being.

Taking stock. I have reviewed the challenge posed to standard welfare economics by behavioral economics. I claimed the challenge is twofold. First, there is a descriptive challenge regarding the assumption that people behave ('as-if') their choices reflect rational preferences. As a consequence, it is claimed that people do not always choose as they prefer. Second, if choices do not always reflect people's preferences, then the implication of this is that using preference-satisfaction as a welfare criterion becomes problematic as well. This is the normative challenge. I then focused on the reply of the well-being approach to these challenges. I have characterized this reply as accepting the descriptive challenge while at the same time rejecting the normative one. This is reflected in two claims of this approach. First, it accepts the idea that the anomalies observed by BE imply that choices do not necessarily reflect individuals' well-being. Second, it takes the path of reconstructing the conception of well-being that would govern people's behavior had they been fully rational. To support the second claim, this approach uses the concept of reasoning failures to explain the less-than-fully-rational behavior of people. We named four such failures: limited technical ability, limited imagination, limited willpower and limited objectivity.

Defined in this way, the well-being approach has two important aspects worthy of further discussion. First, it purports to satisfy only preferences that are not flawed because of mistakes. The use of 'mistakes' here can be elucidated by another term often used in the literature – *purified preferences*.[42] The idea of purified preferences is to consider choices that stem from reasoning failures and deviate from full rationality as mistakes, while decisions that reflect rational behavior as the correct preferences of the individual. Hence, purified preferences can be defined as the preferences individuals would hold in the absence of reasoning failures. Described in this way, the well-being approach purports to satisfy purified preferences and increase people's well-being. Therefore, the satisfaction of purified preferences becomes the constituent of well-being in this approach.[43] One important aspect regarding purified preferences is worth

[42] Strictly speaking, most proponents of this approach do not use the term 'purified preferences' themselves. They use terms such as 'the preferences that would be observed if one was fully rational'. Rather, the term purified preferences is usually used either by critics of this approach or by other scholars in their attempt to explain these ideas. Nonetheless, the term is useful in elucidating the actions and intents of the well-being approach.

[43] For a successful use of purified preferences, the well-being approach needs a way to reconstruct such preferences from choices as well. The attempt to do this is often referred to as preference purification. Some well-known examples of preference purification are: Salant & Ariel Rubinstein (2008); Bordalo,

considering here. How do behavioral paternalists tend to think about them? More specifically, do behavioral paternalists really believe that people have purified preferences that exist in their minds? That some sort of a rational self, free from reasoning failures, must exist, and that it is this self's preferences that must be satisfied? Or do they think about purified preferences in some technical sense, without postulating their existence? That is, they posit the idea of rational preferences as a mere technical tool to help them identify behavior that is free from reasoning failures. For now, we shall put aside these questions. We shall go back to discuss them in Chapter 3. However, it is good to keep them in mind.

The second important aspect of the well-being approach has to do with purified preferences as well. On the one hand, this approach accepts the claim that people can be subject to reasoning failures and consequently make mistakes in their choices. Yet, at the same time, it does not leave people to face their mistakes alone. Rather, this approach aims to help people make better decisions, when better is measured by satisfying their purified preferences instead of their mistaken ones. Thus, the other main aspect of the well-being approach is its interference with people's choices to make them better off. For this reason, this approach (consciously) admits of having some paternalistic aspect to it.

This approach (with some differences) has been called by various names in the literature, notably "*Light Paternalism*",[44] "*Libertarian Paternalism*",[45] "*Asymmetric Paternalism*"[46] and "*New Paternalism*".[47] However, to avoid cumbersome terms, I shall henceforth refer to all these as *Behavioral Paternalism* (BP).[48] This is not an arbitrary choice. The main idea of all these approaches is similar. They have a

Gennaioli, and Shleifer (2013); Bleichrodt, Pinto, and Wakker (2001). This is a non-exhaustive list, of course.

[44] Loewenstein and Haisley, "The Economist as Therapist: Methodological Ramifications of 'Light' Paternalism".

[45] Sunstein and Thaler, "Libertarian Paternalism Is Not an Oxymoron"; Richard H. Thaler and Cass R. Sunstein, "Libertarian Paternalism," *American Economic Review* 93, no. 2 (2003): 175–79.

[46] Camerer et al., "Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism.'"

[47] Roberto Fumagalli, "Decision Sciences and the New Case for Paternalism: Three Welfare-Related Justificatory Challenges," *Social Choice and Welfare* 47, no. 2 (2016): 459–80.

[48] I follow Whitman & Rizzo (2016), in using this term. However, Whitman & Rizzo do not provide a good definition for what is BP. Rather, they define it by examples of policies and ideas that belong to this approach.

paternalistic aspect that purports to influence individuals' behavior to make them better off; and they justify the interventions by pointing to behavioral insights about people's reasoning failures. In other words, in contrast to other forms of paternalism that can be coercive or invoke an objective notion of the good, BP claims to improve people's well-being by referring to their own judgments reflected by their purified preferences. Hence, I believe that the term that reflects best this approach is Behavioral Paternalism.

Let us illustrate what BP is by discussing some prominent examples of it. First, consider Thaler & Sunstein (2008)[49]. Their proposal is to "nudge' (i.e., influence) people's behavior in order to increase their well-being. Formally, the idea is to 'make choosers better off, *as judged by themselves*'.[50] Thus, the welfare analyst (in their terms a 'choice architect') uses people's own judgment about what is good for them as a welfare criterion. Thus, the goals or ends of people's lives are left for them to determine. Yet, Thaler & Sunstein also allow for mistakes in the way people pursue these goals, hence, arguing that intervention in choices can improve their well-being. In other words, the analyst respects people's goals, but still takes into account that people can make mistakes in achieving these goals, or at least err in finding the best means to achieve them. A second and similar approach is Camerer *et al*. (2003)[51]. They argue that a welfare policy is justified if the benefits from it for the irrational individuals are greater than the total costs for the rational individuals as well as other affected parties (e.g., businesses). This approach is similar to the "nudge" policy, because it accepts that individuals can make errors in their choices due to reasoning failures as well. Hence, both approaches claim that people can benefit from interventions in their choices.

Another example of BP is Bernheim (2016). [52] His suggestion is to *limit the welfare domain* by removing inconsistent choices from the more general welfare domain. In other words, the goal is to try to identify a restricted welfare domain that excludes preferences and choices based on errors. Then, as a second step, the analyst can conduct a standard welfare analysis on the restricted domain using the revealed preference

[49] Thaler and Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*.
[50] Ibid., 5, Italics in the original.
[51] Camerer et al., "Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism.'"
[52] B. Douglas Bernheim, "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics," *Journal of Benefit-Cost Analysis* 7, no. 1 (2016): 12–68.

method. One notable aspect of this approach is that in the first step non-choice data can be used to determine the limited welfare domain. Similar ideas exist in Bernheim and Rangel's (2007, 2009) Generalized Choice Situation.[53]

A further prominent example is *Informed Decision Utility*. Here the suggestion is to ensure that the individual was truly informed before making the choice. The assumption is that the decisions that are more informed lead to less reasoning failures. Note that this approach presumably purports to improve the quality of the decision by providing information instead of actively influencing choices. However, because the way of presenting information can frame a decision as well, this approach encounters a problem. Using it can require another criterion of how to provide information in a way that will unproblematically frame a choice. Or put differently, it needs a criterion of how to frame the choice problem such that people would choose according to their purified preferences. Thus, it requires an operational framework in-itself, the lack of which was the original problem.[54]

Lastly, consider Lowenstein & Haisley's (2007) *Pragmatic approach*.[55] They claim that BP policies should only be used when welfare judgments are presumably straight forward. Formally, it entails meeting one of the following three conditions. First, when a decision failure is dominant. For example, when employees do not take advantage of retirement accounts offered by their workplaces.[56] Second, when there are clear negative outcomes. That is, when it is clear the individual would be better had he avoided certain outcomes. For example, it is claimed that it is more probable for people who have a payday loan to go bankrupt. Since (it is assumed that) bankruptcy is a clearly negative outcome, then alternative credit products (i.e., other than payday loans) would be welfare improving for many people. Third, when choices are self-officiating. That is, letting people choose their own goals as long it is not done in the 'heat of the moment'.

[53] B. Douglas Bernheim and Antonio Rangle, "Choice-Theoretic Foundations for Behavioral Welfare Economics," *American Economic Review* 97, no. 2 (2011): 464–70; B. Douglas Bernheim and Antonio Rangle, "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics," *The Quarterly Journal of Economics* 124, no. 1 (2009): 51–104.

[54] See Loewenstein & Haisley (2007, pp. 16-18) for more details about the problems with this approach.

[55] Loewenstein and Haisley.

[56] This further assumes that the employee prefers more income over less and that the employer has a matching retirement program to the employee contribution.

One clarification is perhaps needed. It might be objected that these approaches have some differences between them. They do. Nevertheless, I believe these are minor differences. Their main idea, however, is the same. To repeat, they have a paternalistic aspect that purports to influence individuals' behavior to make them better off; and they justify the interventions by pointing to behavioral insights about people's reasoning failures. But to be explicit, in this discussion I am not evaluating only a certain axiom of rationality or a specific behavioral paternalistic interference. Instead, I try to examine what can perhaps be called the paradigmatic types of interferences used by BP to increase people's well-being, when considered in isolation. That is, behavioral paternalistic policies that purport to increase each person's subjective well-being. Thus, we leave out of the analysis questions about social well-being. We also leave out cases where behavioral policies are used to achieve some public policy objective and not improve well-being. But having made these qualifications, this discussion should apply to all types of behavioral paternalistic interventions that rely on the notion of satisfying purified preferences as a welfare criterion.

Conclusion

Standard welfare economics uses the satisfaction of preferences as a welfare criterion. But since preferences are assumed to be reflected in choices, they should take into account everything that the agent judges to be relevant for the decision. As such, the axioms of rational choice become justified as properties of sound reasoning about choice. However, economists have had different interpretations for the question of 'why preference-satisfaction matters for welfare'. As long as preferences were assumed to comply with rationality, preference-satisfaction could be used as a welfare criterion and welfare economics could remain silent on these interpretations. Yet, evidence from BE has been challenging the assumption about rational behavior. Consequently, this has both descriptive and normative implications. On the descriptive level, the assumption that people display behavior that complies with rationality is challenged. This leads to the claim that people do not always choose as they prefer. Second, if choices do not always reflect people's preferences, then using preference-satisfaction as a welfare criterion becomes problematic as well. I called these the *descriptive* and *normative* challenges, respectively. I then discussed different responses for these challenges, focusing on the well-being approach, which I called *Behavioral Paternalism*. I have characterized BP as an approach that accepts the

descriptive challenge while at the same time rejecting the normative one. I have then discussed two important aspects of BP. First, its welfare criterion is the satisfaction of purified preferences. Second, because of its intervention with choices, BP has a paternalistic aspect. In the next chapter, we shall discuss to greater detail what paternalism exactly means in this discussion.

Chapter 2: Defining paternalism

In the previous chapter, we characterized BP as an approach that accepts the descriptive challenge and rejects the normative one. In its attempt to reconcile the welfare criterion with the descriptive challenge, BP uses the satisfaction of purified preferences. We also saw that by accepting the claim that people make mistakes and by purporting to satisfy only purified preferences, BP has a clear paternalistic aspect. Its intent is to increase people's well-being by interfering with their choices.

However, paternalism itself is a notion often discussed in political philosophy. Therefore, one might wonder if we should not clarify it a bit more in this discussion. This is important because the literature on paternalism is quite dispersed to the extent that different people often use the concept to convey different things. Thus, it was even claimed that one should only use the term with its appropriate adjectival modifiers (e.g., soft-paternalism, means-paternalism, broad-paternalism).[57] But to motivate this chapter even further, this inquiry is not important solely for the purpose of trying to be more precise in using the term 'paternalism'. Engaging with definitional issues of paternalism is also important for the argument this thesis is trying to make. As we shall see, one of the outcomes of this inquiry will be that BP is a means means-paternalistic account. This insight will be important in the arguments made in the next chapter.

The question this chapter is concerned with is, 'What kind of paternalism is BP?'. To answer that, we shall first flesh-out more carefully what is paternalism in general and how it is often understood in economics. However, I shall argue that prominent definitions in the literature are unsatisfactory in describing the paternalistic aspect of BP. The reason is that the respective authors of these definitions have had a certain goal in mind. Instead of trying to capture the ideas of behavioral paternalists, they ventured to understand what seems to be an appropriate definition for a paternalistic intervention in economics. This shall be done in Section 1. I then evaluate how BP's proponents perceive their paternalistic intents. Drawing on that, I suggest a definition for paternalism that reflects best BP. This is done in Section 2. The chapter then examines to what cases the definition can be extended. It is argued that there are three cases in

[57] Thaddeus Mason Pope, "Counting the Dragon's Teeth and Claws: The Definition of Hard Paternalism.," *Georgia State University Law Review* 20, no. 3 (2004): 661.

which BP purely scores. Those are: means-paternalism, broad-paternalism and welfare paternalism. This is done in Section 3.

1. What is paternalism?

The definition of paternalism is often traced to Gerald Dworkin's (1972) well-known work on the topic. Dworkin understands paternalism as "the interference with a person's liberty of action justified by reasons referring exclusively to the welfare, good, happiness, needs, interests, or values of the person being coerced." [58]

An example of a paternalistic policy can be the laws requiring people to use seat belts or motorcyclists to wear helmets. However, to some, this definition seems to be a bit too narrow.[59] The reason is that not every paternalistic action seems to restrict liberty, at least not when liberty is narrowly understood as some sort of external interference with people's choices in the form of coercion. For instance, when a doctor lies to his terminally ill patient by telling that his son, who died in a car accident, is still alive, he acts paternalistically since he believes that lying to the patient will make him happier. But it is hard to see how the doctor's actions coerce or infringe upon the patient's liberty. Thus, cases of the latter sort seem to have a broader sort of interference with people's lives than simply restricting their liberty. Rather, such cases commonly thought of as interference with their autonomy.[60] Therefore, we might say that a more accurate definition of paternalism has to do with some interference with either liberty or autonomy of the person so constrained by the interference, for the sake of making him or her better off.

Yet, even this more broad definition can encounter problems. In particular, the current discussion on paternalism remains quite divided regarding the definition of the term with many caveats and qualifications still being raised.[61] Examples of issues being voiced concern questions of whether paternalism has to do with the intent of the intervening person? Or how should the definition be extended? That is, should the

[58] Gerald Dworkin, "Paternalism," *The Monist* 56, no. 1 (1972): 64–84.
[59] Bernard Gert and Charles M. Culver, "Paternalistic Behavior," *Philosophy & Public Affairs* 6, no. 1 (1976): 45–57; Bernard Gert and Charles M Culver, "The Justification of Paternalism," *Ethics* 89, no. 2 (1979): 199–210.
[60] Gerald Dworkin, "Paternalism: Some Second Thoughts," in *The Theory and Practice of Autonomy* (Cambridge University Press, 1988), 121–29.
[61] David J Garren, "Recent Work," *Philosophical Books* 43, no. 1 (2002): 5–22.

interference be of means or ends? Direct or indirect? Pure or impure? (terms we shall go back to later in this discussion).

But what about paternalism in economics? How is paternalism in economics generally understood? There is no straightforward answer to this question, as various definitions have been suggested for paternalism in economics. Let us review some of the best exponents of these definitions.[62]

First, Le Grand & New (2015)[63] suggest a definition of paternalism that is based on behavioral insights. According to them, an "intervention is paternalistic with respect to an individual if it is intended (a) to address a failure of judgment by that individual and (b) to further the individual's own good".[64] By failure of judgment, they mean the reasoning failures we have reviewed in the former chapter. Another similar definition of paternalism is Fumagalli (2016).[65] According to him, an intervention is paternalistic if it a) violate (or interfere with) the autonomy of its target agents; b) is implemented without the explicit consent of the agent; and c) is designed with the primary aim to enhance the well-being of the agent. Consider also Haybron & Alexandrova (2011).[66] According to them, paternalism by A toward B's behavior (whether through action or through omission of a choice) occurs when: a) the intervention is aimed to have (or to avoid) an effect on B or her sphere of legitimate agency; b) that involves the substitution of some other entity's judgment or agency for B's; c) is done (or omitted) for the sake of B's own interests or matters that legitimately lie within B's control; and d) manifests a non-deferential attitude to B's judgment or agency with respect to those interests or matters. 'Non-deferential attitude' in this context, can mean that A refuses to acknowledge B's agency or authority to make the ultimate decision, even when granting B's wish. Lastly, consider Hausman (2018).[67] According to him, "[a] policy is paternalist with respect to some agent if and only if it aims, for the benefit of the agent,

[62] Some of the following definitions are developed for government interventions only. However, arguably, they can easily be extended to cases in which other private institutions and agents act paternalistically, such as doctors that lie to their patients. Thus, evaluating them is still relevant.
[63] Le Grand and New, *Government Paternalism: Nanny State or Helpful Friend?*, 2015.
[64] IBid.
[65] Roberto Fumagalli, "Decision Sciences and the New Case for Paternalism: Three Welfare-Related Justificatory Challenges," *Social Choice and Welfare* 47, no. 2 (2016): 459–80.
[66] Haybron and Alexandrova, "Paternalism in Economics".
[67] Daniel M. Hausman, "Efficacious and Ethical Public Paternalism," *Review of Behavioral Economics* 5, no. 3–4 (2018): 261–80.

to substitute the policy-maker's determination of what the agent should do for what is properly within the agent's own legitimate domain of judgment or action".[68]

Are these definitions suitable for describing BP? Arguably, no. First, Le Grand & New's definition does not specify how the 'individual's own good' is measured. For example, is it his own subjective good or perhaps some objective measurement such as income or education? A crucial component of BP is that it accepts individuals' own judgments about what is good for them. That is, their subjective well-being. Furthermore, their definition does not say anything about the autonomy or liberty of the agent, and as we shall see, respecting the agent's liberty is a crucial part of BP's paternalism. Fumagalli's definition seems slightly more specified. However, it does not specify the reason for the paternalistic intervention. That is, how is the intervention justified? (e.g., by pointing out to reasoning failures of the agent or something else). Haybron & Alexandrova's definition seems unsatisfactory for similar reasons. First, it, too, lacks the justification for the intervention. Second, as we shall see later, one of BP's main paternalistic aspects is that it is a means-paternalistic approach. That is, a paternalistic account that respects people's choice of ends. Thus, it is not clear at all that BP refuses to acknowledge one's agency or authority to make the ultimate decision. Lastly, Hausman's definition seems too broad and lacks some crucial components to describe BP's paternalism adequately as well. First, it does not specify what does 'benefit' means (again, from the agent's subjective perspective or some objective one?). Second, it lacks a justification for the intervention as well.

I believe that a possible reason why these definitions are unsatisfactory in describing BP is that their respective authors had a certain goal in mind. Instead of trying to capture the ideas of behavioral paternalists, they ventured to understand what seems to be an appropriate definition for a paternalistic intervention in economics. Thus, although these definitions can have merit, I believe that for the purpose of this discussion they are unsatisfactory. To be sure, discussing them was not in vain. In the next section, I shall suggest a definition for BP's paternalism. However, before one suggests a definition for something, it is more appropriate to evaluate what has already been done in the literature and venture from there. In any case, because these definitions are

[68] Hausman, 265.

unsatisfactory for this discussion, a better strategy, I suggest, is to evaluate how BP's own proponents understand their paternalistic actions. In the next section, we shall do just that.

2. A definition for BP's paternalism

Let us examine how some of the leading behavioral paternalists understand their own paternalistic actions. According to Thaler & Sunstein, "a policy counts as 'paternalistic' if it is selected with the goal of influencing the choices of affected parties in a way that will make those parties better off".[69] Thaler & Sunstein (2008), add to this definition the important clause of "*as judged by themselves*".[70] Loewenstein *et al*. (2007),[71] denote that, "paternalistic policies have the goal of benefiting people on an individual basis, premised on the idea that people cannot be relied upon to pursue self-interest". In a like manner, Camerer *et al.* (2003) indicate that "a regulation is…paternalistic if it creates large benefits for those who make errors, while imposing little or no harm on those who are fully rational".[72] Thaler (2015) denotes that: "[b]y paternalism, we mean trying to help people achieve their own goals… we have no interest in telling people what to do".[73] Similarly, Sunstein (2018), claims that the lodestar of behavioral paternalism is people's own judgments.[74] Lastly, Bernheim (2016), asserts that behavioral economics shows that people err when choosing the alternative that leads to their ultimate objective. But behavioral economics does not provide us with evidence that people make mistakes about such ultimate objectives.[75]

Before trying to derive the components of a paternalistic account from the statements above, an example might be helpful. Consider a well-known issue of saving for retirement, an example often used by behavioral paternalists. In the U.S, there seems to

[69] Richard H. Thaler and Cass R. Sunstein, "Libertarian Paternalism," *American Economic Review* 93, no. 2 (2003): 175; Sunstein and Thaler, "Libertarian Paternalism Is Not an Oxymoron," 1162.
[70] Thaler and Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Italics in the original.
[71] Loewenstein and Haisley, "The Economist as Therapist: Methodological Ramifications of 'Light' Paternalism," 5.
[72] Camerer et al., "Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism,'" 1212.
[73] Richard H. Thaler, *Misbehaving: The Making of Behavioral Economics* (New York: W.W Norton & Company, Inc., 2015), 324–25.
[74] Cass R. Sunstein, "'Better off, as Judged by Themselves': A Comment on Evaluating Nudges," *International Review of Economics*, 2018, 2.
[75] Bernheim, "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics," 17.

be presumably good evidence that people under-save. This is substantiated both by a low aggregate saving rate and individual self-reports in which people testify that they have less than they would like.[76] This issue becomes puzzling when considering that employees can voluntarily join a savings program offering them tax benefits and often matching employers' contributions by simply filling a form. Often, behavioral paternalists explain this by pointing out to biases coupled with reasoning failures. First, employees might suffer from the status quo bias. If the default is set out so that they are not registered to the program, then behavioral science predicts that they are likely to preserve in that default – i.e., in their choice not to join. Second, reasoning failures such as limited technical ability can make it difficult to take-in available information about the saving plan and reinforce the choice not to join. For example, perhaps the employee briefly heard about the saving plan in an orientation meeting on his first day at work and was told that if he would like to join, then he needs to go to some office at the other side of the building and fill a form. That can arguably make him forget about it by the end of the day. A notable behavioral policy in this area is to change the status quo so people would sign up for the program by default and opt-out should they choose to do so.

Taken together, four components of a paternalistic account are reflected in the example and statements above. First, there is some interference with the choices of the person so constrained by the interference. In the example of saving for retirement, it is quite straightforward that the choice to change the default is interfering with people's choices. The aim of the policy is to interfere with the choice not to join the program. Second, the interference is justified by some reasoning failure(s). As the saving example shows, the change of the default is justified by referring to reasoning failures and biases such as limited cognitive ability and the status quo bias. Third, the interference is done for the sake of the person's own good. If people admit not to save as much as they would like, they would surely claim to be better off if their goal to save more is satisfied. Fourth, there is no infringement upon liberty, narrowly understood as coercion.

[76] B. Douglas Bernheim, "Do Households Appreciate Their Financial Vulnerabilities? An Analysis of Actions, Perceptions, and Public Policy," in *AM. COUNCIL FOR CAPITAL FORMATION, TAX POLICY FOR ECONOMIC GROWTH IN THE 1990*, 1994, 1–30; Steve Farkas and Miles Johnson, "MILES TO GO: A STATUS REPORT ON AMERICANS' PLANS FOR RETIREMENT 9," 1997.

Changing the default arguably does not entail any sort of such infringement. If people want to, they can always opt-out.

However, three clarifications are immediately required for this definition. First, we need to make explicit what is meant by 'people's own good'. For behavioral paternalists, the aim of the intervention is important. BP policies want to influence people to achieve goals that will make them better off. If the government encourages employees to join a saving program solely for the reason of preventing the need to spend money on them later on by supporting them when they retire, then the government is interfering with choices but not acting paternalistically in the relevant sense here. This is because the aim of the interference is not to make people better off.

Second, the emphasis on people's own good also raises the question of 'what is people's own good?' I argue that BP understands people's own good as their well-being. As we have seen in the previous chapter, BP is a well-being approach. Thus, it only seems reasonable for behavioral paternalists to justify their interventions by appealing to people's own good, understood as their well-being. But if BP interprets people's own good as their well-being, then it seems reasonable that the sort of well-being we should be thinking of here is subjective well-being. To reinforce this point, recall Thaler & Sunstein's "as judged by themselves" clause, or Sunstein's claim that the lodestar is people's own judgments. Also, recall Bernheim's claim that people do not err regarding their ultimate objectives. How should we understand all this if not as a subjective judgment of the agent regarding what is good for him?

The third clarification is about autonomy. The last component of the definition states that there is no infringement upon liberty, narrowly understood. But what about concerns that are broader than simply blocking choices or coercing people? What about potential infringement upon autonomy? We saw in Section 1 of this chapter that the general discussion about paternalism tends to take in considerations of autonomy and not just liberty. In the case of BP, the idea that some people need help to make better decisions (even if it increases their well-being) can be perceived as treating them like little children who are incapable of managing their own lives.[77] Thus, some have argued

[77] Le Grand and New, *Government Paternalism: Nanny State or Helpful Friend?*, 2015, chap. 6.

that BP's policies are paternalistic but for a different reason than interfering with choices for people's own good.[78] The alternative claim is that BP policies are paternalistic because they push individuals to make one choice rather than another by using biases and reasoning failures, which is a potential infringement upon their autonomy. To be sure, interfering with choices for one's own good would still be part of the definition for paternalism in this alternative view. But the emphasis is on the way or method by which BP operates. The claim is that instead of attempting to persuade people by facts or direct arguments, BP's proponents take advantage of reasoning failures and biases to shape people's choices and that this is what makes it paternalistic. Unfortunately, as important as this alternative view is, I will not discuss it elaboratively here, and for a good reason. The focus of this discussion is on the well-being aspect of BP. Thus, the relationship between BP and its ability to do what it claims to – increase people's subjective well-being – should be at the center of the inquiry.

To be sure, it is not as if the two topics are completely unrelated. If the arguments from autonomy are convincing, then treating people like children can, in principle, affect their well-being.[79] For example, some psychological theories claim that BP's interferences with choices can decrease well-being by curtailing agents' motivation to change their behavior. The claim is that if you treat people like children, then they might turn into ones, thus be less capable of making better decisions on their own. But since my aim in this discussion is to evaluate BP's defensibility on its own grounds, it is worth considering how central the issue of autonomy is for BP. From that perspective, behavioral paternalists do not seem too concerned about issues of autonomy. They simply weigh them. The idea is that for some people, some autonomy would be violated. Yet, this violation can be very minimal. Furthermore, for behavioral paternalists, violations of autonomy should not necessarily lead to a decrease in well-being. Sometimes the two can be traded-off. Therefore, unless we are willing to make autonomy a right that can never be violated, then there is still wiggle room to argue in favor of BP. In particular, where we can judge that a policy would lead to large gains in well-being with minor violations of autonomy, then BP can perhaps be justified. In

[78] Daniel M. Hausman and Brynn Welch, "Debate: To Nudge or Not to Nudge," *Journal of Political Philosophy* 18, no. 1 (2010): 123–36.
[79] I rely in this part on Le Grand and New, *Government Paternalism: Nanny State or Helpful Friend?*, 2015, chap. 6.

any case, I shall not pursue this discussion about autonomy anymore. Instead, I shall assume that violations of autonomy might be caused by some (if not all) BP policies, but that such policies do not always have to decrease well-being. What is left to do, and what I shall do in the next chapter is to evaluate if BP policies can actually increase subjective well-being. That is, on people's own judgment.

Let us take stock. I have reviewed definitions for paternalism in economics and claimed that they seem unsatisfactory in describing BP. I then suggested a new definition for BP's paternalism. This definition has four components. First, there is some intervention with choices. Second, the intervention is justified by pointing out to reasoning failures. Third, the interference is done for the sake of the agent's own good, understood as his subjective well-being. If the intervention increases well-being as a simple side effect and was not intended to do so, it will not be considered as paternalistic according to BP. Fourth, the intervention does not infringe upon liberty, narrowly understood as coercion. If it infringes upon autonomy, the infringement is weighted by comparing the possible gains in well-being. In the following section, we shall examine to what cases this definition should be extended.

3. Extensions of the definition

Having suggested a definition for behavioral paternalism, the next natural step is to examine to which cases the definition should be extended. Traditionally, definitions of paternalism are discussed in five different areas of possible extensions. Let us now review these and evaluate how BP scores on them.

2.1 Soft vs hard

Soft paternalism is the view that paternalistic actions are only justified when applied to cases where people act in a non-voluntary way.[80] A non-voluntary action is one where people are either insufficiently informed or capable of rational self-governance.[81] Soft paternalistic interventions are also justified where there is also only the suspicion that a person is not acting voluntarily. However, in such cases of mere suspicion, if during the intervention the intervening agent has determined that the action taken was voluntary, then the paternalistic intervention should be removed. In contrast, a hard

[80] Joel Feinberg, *The Moral Limits of the Criminal Law Volume 3 : Harm to Self*, vol. 3, 1989.
[81] Alex Voorhoeve, "A Response to Rabin," in *Behavioural Public Policy*, 2012.

paternalistic intervention claims to make people better off, even if it was acknowledged that they acted voluntarily.

This distinction is often illustrated by Mill's example of a person walking across a damaged bridge.[82] Mill asks us to imagine a person that is about to cross an unsafe bridge. If we cannot warn him (perhaps there is not enough time before he reaches the dangerous spot where the bridge would collapse or we do not speak his language), then we can seize him by force to remove him from the bridge. Such action would not count as paternalistic. The assumption is that if the person knew the rickety situation of the bridge, he probably would not have wanted to make the choice to cross it. Thus, the choice to cross the bridge was not really his. However, suppose that we were able to stop the person and explain to him that he is about to cross a damaged bridge and that if he proceeds in his actions there is a likely risk for him to fall, injure himself and perhaps even die. If the person replies that he is aware of the situation and that he actually tries to commit suicide, a soft paternalistic approach will entail that we should leave him alone to cross the bridge. In contrast, a hard paternalistic policy would stop the person even after his reply that he wants to commit suicide. This is because the intervening party believes that 'not committing suicide' is better for that person's well-being, whether he admits it or not.

Are BP policies soft or hard paternalistic? This is a hard question to answer. At first glance, it seems that many cases in which BP calls for intervention are cases of soft paternalism.[83] For example, the idea that people under save for their retirement is substantiated, as we have seen, by the fact that they are not aware of the savings program or did not fully take in the available information. Other policies, such as binding commitment that allow gamblers to lock themselves out of casinos by putting their names on a list, can be justified by insufficiency for rational self-governance.

However, arguably there are situations in which people act voluntarily, yet BP still might call for paternalistic interventions. For instance, suppose John wants to lose weight. Yet, he quite often eats high-calorie foods simply because he does not have the

[82] John Stuart Mill, *On Liberty* (London: John W. Parker and Son, West Strand., 1859). See for example Dworkin (2019) and Pope (2014) who use this example in explaining soft vs hard paternalism.
[83] Indeed some explicitly refer to BP as soft paternalism (cf. McQuillin and Sugden, 2012; Sugden, 2008).

proper means to know what is inside that food. It is not as if he is not able of rational self-governance. Nor is he not informed that high-calorie food can undermine his goal. Thus, arguably, he seems to act voluntarily. Rather, what seems to be happening here is that John lacks the means to achieve his goal. A simple, and often advocated, BP policy of labeling clearly the calories in foods, can help John to achieve his goal. Thus, it seems that some choice situations are cases in which people can be considered to act voluntarily, yet they still do not choose the best means to achieve their own ends. I admit, this is a bit of a crude distinction. It is perhaps possible to extend soft paternalism also to cases where people do not have the proper means to achieve their ends. One can argue that a rational person who wants to lose weight would take the effort to find out the calories in the foods he consumes, just as a rational person would perhaps find out if the bridge he is about to cross looks safe. Yet, I think that extending soft paternalism in this way is unnecessary. The main reason is that the following extension of paternalism already accommodates that.

2.2 Means vs Ends

In means-paternalism, the intervening agent respects other people's end or goals, yet at the same time accepts that they can err in trying to achieve these ends.[84] Thus, the interference is extended only to the means that people choose to achieve their ends. Ends-paternalism, on the other hand, would claim that people can make mistakes in choosing their ends and therefore interventions should be extended to people's ends as well. For example, suppose we see a person driving a motorcycle without a helmet. If he says that safety is important to him, then a means paternalistic approach would require him to wear a helmet to reduce the risk of accident and thus satisfy his end for safety. However, suppose that instead he claims that his goal by driving without a helmet is feeling the wind blowing through his hair. In that case, the requirement to wear a helmet would be considered as ends-paternalistic.[85]

I argue that BP is solely means-paternalistic. To have an ends-paternalistic claim, BP would need to argue that people make mistakes in their goals; that some of their

[84] Thoma, "Merely Means Paternalist ? Prospect Theory and ' Debiased ' Welfare Analysis"; Julian Le Grand and Bill New, *Government Paternalism: Nanny State or Helpful Friend?* (Princeton University Press, 2015).

[85] These examples are taken from Gerald Dworkin, "Paternalism," in *The Stanford Encyclopedia of Philosophy*.

conceptions of the good are false. This would be equivalent to telling people that the desire or satisfaction they feel from engaging in a certain activity is false. Consider the motorcyclist again. If his goal is to enjoy the wind in his air, then claiming he is mistaken is like telling him that the sensations he feels when driving without a helmet are illusions. On the other hand, a means-paternalistic approach might simply ensure that he understood the risks he is about to take, whether by putting vivid road signs or requiring some training about these risks (e.g., taking a mandatory driving course before he can get his license). Since BP is a subjective account of well-being, it is hard to see how it can make judgments about people's conceptions of the good – i.e., about their ends. Doing so would arguably undermine one of its core aspects.

However, this might be too quick. People can often have ends that conflict with each other. For instance, the motorcyclist can both hold the goal of enjoying a drive without a helmet to feel the wind in his hair, and at the same time, the goal of living a long healthy life. Yet, when deliberating if he should wear the helmet or not, he might prioritize his momentary enjoyment of feeling the wind in his hair over the goal of living a long healthy life. Thus, it seems possible that people often weigh and balance their ends at different moments in time. But if this is the case, then it seems that some mistakes of means can relate to choices of ends as well, thus, obfuscating the means-ends distinction to some extent. For example, how can we know that the motorcyclist weighed his two ends correctly? Is it possible that due to limited technical ability he does not compute correctly how his preference for 'wind in the hair' can undermine his 'long healthy life' goal? It seems then that the claim that BP is purely means-paternalistic needs to be qualified a bit. As a means-paternalist account, behavioral paternalists need to accept the balance of ends people choose for their lives, *unless there is a reasoning failure in choosing that balance*.[86] That is, when judging the appropriateness of balancing ends, behavioral paternalists look for reasoning failures just as they do in evaluating people's choice of means. To be clear, this is still different from embracing an ends-paternalistic account. An ends-paternalist account would evaluate or question people's ends, not the way these ends are weighted. Of course, the outcome might be the same. The motorcyclist might be interfered with his choice so

[86]See Le Grand and New (2015) for a more elborative account of this qualification.

that he would wear the helmet. Yet, the justification for the outcome would different, and that is important here.

2.3 Broad vs Narrow

Narrow paternalistic interventions are concerned only with interventions by the government. The laws requiring people to use seatbelts are an example of a narrow intervention. Broad interventions are concerned with any party intervening with choices (including the government). For instance, doctors who lie to their patients to make them better off are examples of a broad paternalistic intervention.[87] It is easy to see that BP is broad paternalistic since behavioral paternalistic policies are used by both governments and other private institutions and agents.

2.4 Pure vs Impure

In cases of pure paternalism, the group interfered with is equal to the group protected by the policy. For example, some countries require their citizens to take medical insurance. This is done for the sake of the citizens good, but it also interferes with their choice (they cannot choose not to take an insurance). In cases of impure paternalism, the two groups are not equal. An example of impure paternalism is the warning signs on cigarettes. This intervention is interfering with the cigarette manufacturers for the sake of the consumers (some manufacturers might be smokers, of course, but the idea is that the two groups are not necessarily the same).[88]

BP's policies can be both pure and impure. For example, the savings program for retirement discussed in the previous section is a pure paternalistic policy. It interferes with people's choices to protect them. An example of an impure BP policy would be the cigarette policy mentioned above or special food labeling. Some countries require food labels that are more accessible to the consumer, by using vivid labels to mark high-calorie products or stating how much calories are per food item and just per the entire pack. These paternalistic policies interfere with the manufacturers for the sake of the consumers good and are thus impure paternalistic interventions.

[87] Dworkin, "Paternalism," 2019.
[88] Dworkin, "Paternalism," 1972, 68.

2.5 Moral vs Welfare

Cases of moral paternalism are interferences with choices justified by concerns to people's moral character. That is, to prevent people from leading a morally degrading life. In contrast, welfare paternalism would limit itself to concerns for the person's own welfare.[89] For example, if a State prohibits prostitution, claiming that it is a morally corrupting act, then it uses a moral paternalistic policy. The behavioral paternalistic policies we are concerned with in this discussion are solely cases of welfare paternalism. They aim to increase people's own welfare (or well-being, which is synonymous in BP's case). To be sure, there are behavioral policies that do not target the individual's welfare, but rather some public policy goals. Examples of this are behavioral policies that aim to reduce tax evasion or pollution.[90] Where the intent beyond these goals is to improve people's moral character as well, such interventions can be considered as moral paternalism. In any case, behavioral paternalism, as defined here, is not concerned with moral issues but with issues of welfare.

Taking stock. I have discussed the possible extensions of BP. It seems there are three cases in which BP purely scores. Those are: means-paternalism, broad-paternalism and welfare paternalism. In all the other cases, there are instances in which it can be both soft or hard and pure or impure. However, as we shall see in the next chapter, the means-paternalistic aspect is the most important one for this discussion as it relates to BP's other salient characteristic – i.e., satisfying purified preferences as a welfare criterion.

Conclusion

The literature on paternalism is quite diverse. There is no one definition for the term. Even when applied to economics, there are a plethora of definitions. I suggested a new definition for behavioral paternalism that encompasses the following components: a policy is behavioral paternalistic if a) it intervenes with choices; b) justified by reasoning failures; c) is done to make someone better off; and d) without infringing upon his liberty, narrowly understood. I then reviewed to which cases this definition can be extended. There are three cases in which BP purely scores. Those are: means-paternalism, broad-paternalism and welfare-paternalism.

[89] Gerald Dworkin, "Moral Paternalism," *Law and Philosophy* 24, no. 3 (2005): 305–19.
[90] For examples, see Laure Kuhfuss et al., "Nudges, Social Norms, and Permanence in Agri-Environmental Schemes," *Land Economics* 92, no. 4 (2016): 641–55; Thaler and Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, 67–79.

Chapter 3: connecting the dots. Can BP be defended?

Let us now connect the discussion of the previous chapters. It seems that two characteristics are salient in BP. First, it is a purified preference-satisfaction account of well-being. More specifically, BP is a subjective well-being account in which the satisfaction of purified preferences defines well-being. The second is that it is a paternalistic account. This is because it purports to improve people's decisions by interfering with their choices. Yet, it does so not by imposing goals on their lives but rather by helping them to find the best means to achieve their own chosen ends. Thus, it is a means-paternalistic account.

By purporting to satisfy only purified preferences and claiming to be means-paternalistic at the same time, it follows that BP equates purified preferences with people's own goals. Therefore, if we wish to evaluate whether behavioral paternalists can interfere with people's choices to make them better off by their own standard, then we need to examine whether it is possible to establish the claim that the satisfaction of purified preferences can really define well-being; that satisfying purified preferences is equal to helping people achieve their own goals. In this chapter, we shall evaluate this. I shall argue that, yes, that satisfying purified preferences can be equal to helping people achieve their own goals. However, the outcome may not always be perfect. Sometimes we may need to use safeguards to avoid cases that can curtail people's well-being as well.

To give an overview of the argument: it seems that two main obstacles stand in the way to defend the satisfaction of purified preferences. First, it is contended that behavioral paternalists seem (perhaps implicitly) to think that purified preferences really exist in people's minds. However, such a model of the human mind is argued to be counter-intuitive to some of our daily experiences and even contradict some psychological theories. This critique shall be explicated in section 1. Second, even if purified preferences do exist, it is often argued that to successfully use them to increase the overall well-being of the agent requires a lot of knowledge. Yet, to have access to ample knowledge about the agent seems unrealistic. This critique shall be explained in section 2. Having presented these critiques, we would then examine if BP can overcome them while still holding its claims intact. I shall first argue that behavioral paternalists can

think about purified preferences as mere 'counterfactual-ifs', and that many of them seem to think about purified preferences in this way. This allows behavioral paternalists to avoid making the claim that purified preferences exist in people's minds. Furthermore, I shall explain why thinking about purified preferences in this way does not necessarily imply that behavioral paternalists need to forsake their means-paternalist aspect. Second, I shall argue that BP does not need to show that it can always take into consideration all possible effects on the agent's well-being. Indeed, this seems to require too much knowledge about the agent's goals and reasoning failures. Instead, I shall argue that what BP needs to show is that the analyst can do, at least, better than the individual and, at least, in most cases. This shall be done in section 3.

1. Purified preferences: true or false?

Let us start with what seems to be the first fundamental critique against the use of purified preferences. We have seen that BP's welfare criterion is that satisfaction of purified preferences and that for BP to be a defensible approach, it needs that the notion of purified preferences is undisputed. However, what happens if purified preferences do not exist? As Infante *et al*. (2016) put it:[91]

> [T]he problem we identified in the literature of behavioural welfare economics […] That problem was to justify the implicit assumption that, for any given individual, there exists some mode of latent reasoning that generates complete and context-independent subjective preferences

Similarly Schnellenbach (2019) asserts that:[92]

> By claiming to make people better off according to their own standards, Sunstein and Thaler assume that there exist so-called "purified" preferences that are consistent and stable in the sense that they do not depend on changing contexts, and that characterize the true interests of an individual, from which she deviates with her short-term choices

Likewise Dold (2018) claims that :[93]

> In general, [preference purification] interprets preferences as expressing individuals' subjective judgments about their welfare; they are not synonymous with choice, yet they are causally connected to it. Preference Purification assumes that preferences

[91] Infante, Lecouteux, and Sugden, "Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioural Welfare Economics," 13.
[92] Jan Schnellenbach, "Evolving Hierarchical Preferences and Behavioral Economic Policies," *Public Choice* 178, no. 1–2 (2019): 32, https://doi.org/10.1007/s11127-018-0607-4.
[93] Malte F. Dold, "Back to Buchanan? Explorations of Welfare and Subjectivism in Behavioral Economics," *Journal of Economic Methodology* 25, no. 2 (2018): 162–63.

> exist… In other words: the rational self must exist in order for the normative theory to work

I cannot improve much upon these critics' language. But the essence of the first main critique against purified preferences seems to be this: if welfare analysts are required to satisfy purified preferences, then we need for such preferences to exist in people's minds. We need some 'rational self' to exist.

Why should we find this critique convincing? Or how do the critics substantiate it? There are at least two claims in its favor. First, it is claimed that some of our daily experiences seem to make it convincing. Consider the following example of purchasing a designed fabric: [94]

> A typical consumer, has no idea how to create the designs which, when he sees, he knows he likes. The essential problem is not that he lacks the skill needed to transform a mental image of a design into some concrete form; it is that he cannot form the mental image itself. It is only after a design has been created and put before him that he has any conception of it as a potential object of choice—as something that he has preferences about […] On the most natural description of what is going on, consumers have preferences only for the designs that have actually been created; those preferences are formed as a mental response to that particular set of designs. It seems entirely possible that such preferences will be conditioned by the sets of designs to which they respond […] A design that appears jarringly strange against the background of one set of comparators may seem interestingly original against another, and boringly conventional against a third

In other words, it seems plausible, at least as far as our experience of some choice situations testifies, that we do not have a purified preference for one thing over another, in the sense that BP allegedly requires. As the above example suggests, when we go to buy a shirt, it seems possible that all we know (at best) is that we want a shirt. We do not know what design, color and perhaps even type (e.g., buttoned, v-collar, singlet) we want until we have arrived at the store and evaluated the options. Furthermore, contextual factors that might seem presumably irrelevant to a rational person, such as the color of the wall or the store's lighting, might affect our preference.

[94] Sugden, "Why Incoherent Preferences Do Not Justify Paternalism," 245.

From BP's side, it is possible to reply that this by itself is not an objection. The fact that irrelevant contextual factors affect our choices is well accepted by BP (this is indeed the descriptive challenge it accepts). Furthermore, it can be granted that perhaps people cannot always state exactly what shirt they want before entering the store and evaluate the options. Yet, what BP can claim is that people can make mistakes such as buying a shirt that looks good at the store (perhaps due to the special lighting) and regretting the purchase when trying it at home. Or that people can spend more money than they planned to. BP policies such as cooling periods are installed exactly for these reasons. Thus, if people end up returning a shirt to the store, then perhaps we can assume that they did not really prefer it. Nonetheless, this reply from BP's side may not be sufficient to counter this critique. The problem is that still, on a conceptual level, purified preferences may not exist in the same way it is claimed they are. That is, in people's minds. This takes us to the second reason that substantiates the critique that purified preferences do not exist.

The second reason is that psychological observations and theories of human behavior do not seem to support the existence of purified preferences. We saw that BP claims that satisfying purified preferences constitutes people's well-being; satisfying other preferences is consequently treated as a mistake. Thus, satisfying other preferences that are not purified preferences does not increase well-being. This, the critics argue, depict a problematic model of human agency, in which an inner-rational agent that has all the rational capabilities we have discussed in Chapter 1, is trapped in an error-prone shell.[95] Thus, it is this inner-rational agent that generates the agent's purified preferences. However, the critics argue, we should be skeptical about this model of the human mind, for two reasons. First, empirical evidence shows that when people judge the overall satisfaction with their lives – something that is usually taken as a synonym for well-being – their current focus of attention affects the weights they give to different aspects of their life. If this observation is cogent, then it gives support to the claim that purified preferences are not something stable, that whatever someone *truly* prefers is affected by contextual factors as well. Consequently, this decreases the support for the inner-rational agent model, or at least makes it less convincing. How would behavioral

[95] Infante, Lecouteux, and Sugden, "Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioural Welfare Economics." Unless stated otherwise, the rest of the discussion in this paragraph relies on Infante *et al*. (2016) as well.

paternalists possibly reply to this? In principle, they can object by claiming that the inner-rational agent model is supported by the psychological dual-system theory of the mind. This theory claims that the human mind has two systems: System 1, which makes fast decisions and responds to heuristics; and System 2, which makes decisions more slowly after closer reflection on the choice situation. Therefore, it is possible to explicate behavioral paternalists' claims by assuming that System 2 resembles the inner-rational agent and System 1 the error-prone shell.[96] Indeed, in their explanation of reasoning failures, some of the leading behavioral paternalists have invoked the use of the dual-system theory.[97] However, although this theory is supported by psychological experiments,[98] it has also been quite criticized in recent years.[99] Thus, it remains contested if the dual-system theory can support the inner-rational agent model after all.

Yet, according to the critics, the problem from the psychological side is perhaps more fundamental than the debate about the dual-system theory. The more serious problem is that rational choice is not self-explanatory.[100] This is the second reason why we should be skeptical about the inner-rational agent model. The claim is that cases in which choices comply with rationality are in need of psychological explanation just as presumable deviations or mistakes. However, the inner-rational agent model seems to ignore this point, the critics claim. Instead of providing us with explanations for rational choice, behavioral paternalists depict human psychology "as a set of forces which affects behavior by interfering with rational choice, but rational choice itself –

[96]See Infante, Lecouteux, and Sugden, 15. for a short review of how behavioral economists endorse the dual-system theory of the mind.
[97] See for example, Thaler and Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, 19–22.
[98] Daniel Kahneman, *Thinking, Fast and Slow* (New York: Farrar, Straus and Giroux, 2011); P. C. Wason and J. ST B.T. Evans, "Dual Processes in Reasoning?," *Cognition* 3, no. 2 (1974): 141–54; Daniel Kahneman, "A Perspective on Judgment and Choice: Mapping Bounded Rationality. American Psychologist, 58(9), 697-720.," *American Psychologist* 58, no. 9 (2003): 697–720.
[99] Jonathan St B.T. Evans and Keith E. Stanovich, "Dual-Process Theories of Higher Cognition: Advancing the Debate," *Perspectives on Psychological Science*, 2013; C. Michael Hall, "Framing Behavioural Approaches to Understanding and Governing Sustainable Tourism Consumption: Beyond Neoliberalism, 'Nudging' and 'Green Growth'?," *Journal of Sustainable Tourism*, 2013.
[100] Whitman and Rizzo, "The Problematic Welfare Standards of Behavioral Paternalism," 423; Infante, Lecouteux, and Sugden, "Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioural Welfare Economics."

represented by the error-free reasoning of the inner agent – is not given any psychological explanation."[101]

This is quite a serious problem. If purified preference cannot be adequately supported, then perhaps we should not have much reason to believe they exist. By 'exist', I mean that they exist in people's minds; that human psychology does work in some way that resembles the inner-rational agent model. If this critique is convincing, then the idea that people have a purified preference for something becomes quite vacuous. How can the welfare analyst, the critics might claim, help people satisfy a preference for something that does not exist?

2. The 'knowledge problem' of purified preferences

But suppose now that purified preferences do exist. How can we know what they are? Or at least how can behavioral paternalists claim they have sufficient understanding of them in a sense that their interventions will indeed improve people's overall well-being? This is the second main critique against purified preferences. To elucidate this critique better, it can be helpful to reflect on F.A Hayek's well-known 'knowledge problem'.[102] Hayek argued that a central planner equipped with all the relevant information about endowments, technologies and people's preferences could, in principle, design a market allocation that would be efficient. Yet, Hayek's critique of this idea was that to assume access to such information is to ignore the main problem. That is, that to have an efficient allocation, or at least an allocation that will be as good as the free market one, the social planner will need a lot of knowledge about all these things. However, this seems to require gathering a lot of information, which is not always possible to achieve (and where it is possible, Hayek was concerned that it might infringe upon people's privacy and liberty too much).

How is Hayek's critique relevant to BP? Some critics have argued that it seems plausible that BP can fall prey to a similar knowledge problem.[103] In order for BP to

[101] Infante, Lecouteux, and Sugden, "Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioural Welfare Economics," 14–15.
[102] Friedrich A. Hayek, "The Use of Knowledge in Society F . A . Hayek," *The American Economic Review*, 1945.
[103] Mario J. Rizzo and Douglas Glen Whitman, "The Knowledge Problem of New Paternalism," *Brigham Young University Law Review*, no. 4 (2009): 910; Mark Pennington, "Paternalism, Behavioural

design policies that will improve people's subjective well-being, the analyst needs to have knowledge about people. In principle, if the analyst would possess all the relevant information about people's goals, reasoning failures and the contexts in which such failures appear, then he could implement a policy that would improve people's subjective well-being, by satisfying their purified preferences. However, just like in Hayek's critique, so is in BP's case to assume we have access to all the relevant information is perhaps to assume the problem away. Furthermore, even if we assume that the analyst has enough information such that he can successfully improve people's well-being in one choice situation, it seems unlikely to assume that he has enough information to counter all choice situations – that is, to know whether one intervention will generate negative effects elsewhere. Pennington (2019) is illustrative of this:[104]

> Even if one assumes that public regulators are able to shift individual decisions in one domain, they are unlikely to have knowledge of whether this will generate countervailing effects elsewhere […] people are subject to multiple biases, which may push their behaviour in conflicting directions. Nudging them in one direction while simultaneously failing to deal with the effects of other biases may actually make it more difficult for a person to pursue their ends as a whole. In addition, people may respond to external self-control efforts by reducing their own attempts to develop effective behavioural strategies. The knowledge required to weigh all of these conflicting influences to arrive at policies that on balance improve rather than worsen the scope for people to achieve their ends is likely to be highly contextual and thus largely inaccessible to a public regulatory agency no matter how 'expert' its employees are in the 'science of self-control'

Taking stock. BP needs to defend purified preferences. However, as we have seen, two main challenges stand in its way. First, it is claimed that purified preferences may not exist. BP's claims seem to be captured by what the critics call 'the inner-rational agent model.' In that model, people have purified preferences in their minds. Consequently, the analyst's task is to overcome the error-prone shell in a satisfactory manner that will allow him to successfully reveal the purified preference. Yet, both our experience and psychological theories challenge this model. Second, even if purified preferences exist,

Economics, Irrationality and State Failure," *European Journal of Political Theory* 18, no. 4 (2019): 565–77. I largely follow Rizzio & Whitman in the exposition of this critique.

[104] Pennington, "Paternalism, Behavioural Economics, Irrationality and State Failure," 573.

to successfully know what they are appears to require quite a lot of information about people. However, it seems unlikely that the analyst will have enough access to such information. But even if he does have access to ample information to influence one choice successfully, he might lack knowledge about how that one choice might create negative effects elsewhere, which can reduce the agent's overall well-being.

In the next section, we shall evaluate how BP can counter these critiques.

3. Responding to the critiques

In this section, I shall evaluate BP's response to the critiques above. To give an overview of the argument, I shall first claim that behavioral paternalists do not need to claim that purified preferences exist in some psychological way that resembles the inner-rational agent model. Instead, behavioral paternalists can think about purified preferences as 'counterfactual-ifs'. I shall also claim that thinking in this way does not necessarily entail that behavioral paternalists need to forsake their means-paternalist aspect. This shall be done in section 3.1. I shall then address the second critique. I shall claim that behavioral paternalists do not need to counter every possible tradeoff regarding the agent's well-being. Rather, they only need to be able to do better than the individual and, at least, most of the time. Furthermore, there are good reasons why we can think that they can do so. Moreover, to counter the possibility of cases in which behavioral paternalists may not do better, we can use safeguards. This shall be explained in section 3.2.

3.1 A reply to the first critique

Let us start with the first critique. If that critique is compelling, then purified preferences seem to become a vacuous concept. More specifically, if purified preferences are not supported by our experiences and psychological theories, then how can behavioral paternalists still use them? This becomes even more problematic when considering the means-paternalistic aspect of BP. For an intervention to be considered means-paternalistic, it arguably requires that agents are able to agree that it was good for them; that it improved their well-being. Otherwise, if people do not admit that they are better off, we might say that the analyst interfered with a goal and not a mean of the agent, hence, making the intervention an ends-paternalistic one. But if purified preferences do not exist, then we may ask, how can behavioral paternalists assert to

satisfy them and claim to be means-paternalistic? Can we satisfy something that does not exist in the agent's mind and claim that by doing so we are improving his subjective well-being?

Arguably, this is not a knockdown argument against BP. Behavioral paternalists do not have to accept the inner-rational agent model as true. Instead, they can claim that purified preferences are just a *'counterfactual-if'*. That is, that they use it as a technical notion that helps to elucidate which preferences *would* be observed if people would not fall prey to biases and reasoning failures. This allows behavioral paternalists to avoid biting the bullet of the claim that purified preferences exist in some psychological way in people's minds.

I believe that this way of thinking seems to reflect better the ideas of some behavioral paternalists. For example, Thaler & Sunstein (2003) claim that:[105]

> [C]onsumers will often lack well-formed preferences, in the sense of preferences that are firmly held and preexist the director's own choices about how to order the relevant items. If the arrangement of the alternatives has a significant effect on the selections the customers make, then their true "preferences" do not formally exist

Similarly, Hausman (2016) argues that:

> [I]t seems to me that behavioral economists can deny that they are committed to any inner agent or to any process whereby they construct context-independent preferences. To suppose that there are truth conditions for the claim that an agent has a true or purified preference […] does not commit the behavioral economist to postulating the existence of an inner agent who is capable of weighing the various considerations bearing on the choice […] and determining an overall preference

Another support can be found in Berg & Gigerenzer's (2010) analysis.[106] They describe behavioral economics as using an "as-if" arguments to support its use of rationality in normative claims. In particular, when evaluating the normative part of BE, they claim that:[107]

[105] Sunstein and Thaler, "Libertarian Paternalism Is Not an Oxymoron," 1164.
[106] Nathan Berg and Gerd Gigerenzer, "AS-IF BEHAVIORAL ECONOMICS: NEOCLASSICAL ECONOMICS IN DISGUISE?," *HISTORY OF ECONOMIC IDEAS* 18, no. 1 (2010): 133–65.
[107] Ibid., 102

> [T]he normative interpretation of deviations as mistakes does not follow from an empirical investigation linking deviations to negative outcomes. The empirical investigation is limited to testing whether behavior conforms to a neoclassical normative ideal

A similar claim is found in Beshears *et al.* (2008). They present a method for identifying mistakes by comparing choices with 'normative preferences', which are preferences that comply with rationality.[108] What these examples seem to suggest is that behavioral paternalists seem to think about purified preferences as a useful tool to explain what is a reasoning-failure-free – i.e., rational – behavior. However, they do not necessarily think about purified preferences as something that exists in people's minds.

To be sure, thinking in this "counterfactual-if" way does not imply that behavioral paternalists need to reject their means-paternalistic aspect. The reason is that people can still claim that they want to be rational, or make decisions that are consistent with their other goals, albeit not having a purified preference that exists in some psychological sense in their minds. Indeed, this seems to be somewhat convincing when reflecting upon our daily experiences. Suppose that you meet a person who believes in God but also claims to be a racist. Call him Bob. You tell Bob that since God created everyone equal, he cannot be both a believer and a racist. There is an inconsistency between his beliefs and, therefore, if he wants to be a rational person, he ought to choose between the two. Although it is possible that some would claim they are willing to live with this contradiction, I believe that most people would appreciate this inconsistency as compelling, and upon reflecting on their choices, decide that one of them has to be a mistake. Arguably, it is somewhat evident in this way that people often appreciate rationality; that requirements of rationality are a plausible constraint for them on their reflections. The idea that behavioral paternalists think about rationality in this way is granted even by the critics:[109]

> The element of paternalism in these [behavioral paternalist] proposals can be made more palatable by suggesting not only that their aim is to increase the welfare of the targeted individuals, but also that welfare is being measured according to those individuals' own judgements, and that the choices that individuals are being nudged

[108] John Beshears et al., "How Are Preferences Revealed?," *Journal of Public Economics* 92, no. 8–9 (2008): 1787–94.
[109] Infante, Lecouteux, and Sugden, "Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioural Welfare Economics," 4.

> away from would be mistakes. These suggestions are often expressed through the idea that nudges help individuals to make what, on reflection, they themselves would recognise as better choices

Importantly, if upon reflection, Bob decides that he does not want to be a racist anymore because his belief in God is more important, then he will probably also admit that he is better off; that his well-being is improved now that this mistake had been revealed. People make such corrections in their beliefs all the time and they are usually happier and more satisfied that they have more coherent views (otherwise, they would probably choose to remain inconsistent). To be sure, I am not claiming that being rational ought to be viewed as a goal in itself (although I am not excluding the option that for some people it might be). Rather, I am more alluding here to the possibility that rationality viewed in this way, especially in regards to its consistency requirement, can support the goals people already have and can even improve their well-being (or at least I am alluding in some sense that rationality is positively related to their well-being). If this is convincing, then people can admit upon reflection that they want to be rational, although not having a purified preference that exists in some psychological way. Or put differently, people can appreciate rationality without having an inner-rational agent in their mind. But if people admit that they are better off, then it makes the intervention a means-paternalistic one.

To illustrate the difference between my suggested reply and the claim that purified preferences really exist in people's minds (call it '*the critics claim'*), consider the motorcyclist again from the previous chapter. To recall, his choice is between wearing a helmet or not. However, he seems to have two conflicting goals: 'to feel the wind in his hair', and 'to live a long healthy life'. On the critics claim, because behavioral paternalists are allegedly committed to the idea that purified preferences exist, then in some sense in the motorcyclist mind we should expect to find a purified preference, say, for a long healthy life. Upon trying to decide whether to wear his helmet or not, it seems to the motorcyclist that he feels an equal attraction to both the need to feel the wind in his hair (i.e., drive without a helmet) and to live a long healthy life (i.e., drive with a helmet). Yet, in reality, this attraction is not equal. The inner-rational agent part of the mind has made all the calculations and weighing processes and knows what the agent truly wants – i.e., what his purified preference is. However, another part of the mind –

i.e., the error-prone shell – makes it harder for the motorcyclist to know what his purified preference is. It is like his mind is in some 'psychological war'. His thoughts are obfuscated. He can think of arguments in favor and against wearing the helmet, and it is hard for him to make a decision. If he would only be able to overcome the error-prone part of his mind, then he would be able to achieve his goal for a long healthy life, which is reflected in his purified preference. Otherwise, his choice would be a mistake. In contrast, on the reply I suggested, when trying to decide whether to wear the helmet or not, the motorcyclist really desires two things. However, his moment of deliberation might look the same: obfuscated thoughts and arguments in favor and against wearing the helmet, resulting in difficulty to make a decision. Crucially, there is no part of his mind that has generated a purified preference and is in some sort of 'psychological battle' with the error-prone shell of his mind. Yet, he could still use the help of a third party to make the decision and he can agree that he is better off after the intervention as well. Let us explicate this claim.

Although the agent may not have a purified preference that exists in his mind, he still struggles to make a decision. However, not only he wishes that he would not struggle in his decision making, but he also wishes to be able to make a decision that would be consistent with other goals he has. Finding what he would prefer had he been fully rational (the 'counterfactual-if'), the analyst can consider one alternative as an error, the other being his *allegedly* purified preference. Since the agent values being rational (recall the example above of Bob, the racist believer), then, of course, he would admit he is better off after the intervention. For example, consider the issue of saving for retirement again. Let us assume that after changing the default, the agent admits that his well-being increased. Now that he sees every month money being deposited in his retirement savings account, he says, 'I am better off.' If his answer is consistent with other goals he has such as his long-run goals (perhaps he desires not to work anymore when he is older or to be able to assist his children financially even after he retires), then it seems plausible to assume that his answer reveals that 'saving for retirement' can be considered as his purified preference. Importantly, this would be true even if the agent would say that he is better off under the opposite case in which the intervention was not implemented. That is, in the case that the 'not saving' option remained the default, and the agent did not join a program. If in that case the agent would reply that he is better off, we could still call this a mistake according to BP. This is because it

does not match his other goals, hence, violating the rationality requirement of internal consistency between preferences.

To conclude so far. The first main critique was that purified preferences do not seem to exist. However, I have claimed that behavioral paternalists can (and seem) to think about purified preferences as a counterfactual-if. Thus, behavioral paternalists can still use purified preferences without assuming they exist. Importantly, thinking about purified preferences as a counterfactual-if does not necessarily undermine BP's ability to be means-paternalistic. This is because people can (and seem to) upon reflection admit that satisfying purified preferences is a good thing for them, without having a purified preference that exists in their mind. The reason for this seems that people can value rationality as something that supports their goals.

3.2 A reply to the second critique

But what about the second critique? That is, that even if we successfully managed to intervene with one choice, it might be difficult to know how it will affect other domains in the agent's life such that his overall well-being is not decreased. Thus, to counter this reply, what BP needs to show is that even if some behavioral paternalistic interventions will cause a decrease in well-being in other domains of the agent's life, his overall well-being will not be reduced.

It is possible to suggest at least two replies from BP's side to this critique. First, BP's use of purified preferences can also be perceived as a tool that, in principle, guarantees that the overall well-being of the agent will not decrease. Indeed, the very idea of internal consistency, for example, or of stability, is that there will be no contradictions or rapid changes in the agent's preferences such that his goals as a whole would be met. To illustrate, consider the motorcyclist example again, viewed from the welfare analyst perspective this time. The analyst may ask, 'Did the agent had other goals besides feeling the wind in his hair?' Will these goals be undermined by his choice to drive without a helmet? 'If so, can we say that taking this one choice to drive without a helmet would affect his ability to pursue his goals as a whole?' Thus, it seems that the same reply from BP for the first critique is, to a large extent, valid for the second one as well. In other words, by definition, satisfying purified preferences should imply that the

agent's goals as a whole – that is, his overall well-being – should not be undermined as well.

However, one may object that in replying in this way I assumed the problem away again. That my reply would be valid only if the analyst would really know everything about the agent goals and reasoning failures. Granted, this is somewhat true. But it leads us to discuss the second reply from BP's side. Arguably, what BP needs to show is *not* that it can always take into consideration all possible trade-offs in the agent's well-being. Indeed, this seems perhaps too much to ask. What BP needs to show is that the analyst can do, at least, better than the individual and, at least, in most cases. By 'better,' I mean both that the analyst can identify better what is a mistake and also that he can better find the optimal (or more superior line of action) having identified the mistake. There are at least four reasons why we should think this is possible for BP to achieve.[110]

First, consider the reasoning failure of limited technical ability. If a decision has to do with complex technicalities, say, computing probabilities, then in some cases the analyst can be more equipped than the individual to make the decision. For example, in the case of saving for retirement, it seems plausible that a lot of people do not have ample knowledge about the different investment options or how to calculate the returns correctly in the face of risk. However, an analyst who is familiar with the savings products in the market and is equipped with the knowledge of how to compute the returns correctly could arguably do a better job than most people in that sense. Default retirement plans that assign people to a savings plan based on their age and then change their plan every few years are examples of a BP policy that uses this insight. The idea is that younger people could benefit more from riskier investments as they have more time until their retirement. Thus, they can afford to lose more money due to the riskier investment, but at the same time could enjoy more the fruits of compound interest on the higher and riskier gains as well. The older people become, however, they might be better off in a more conservative investment plan as they have more to lose and less time to accumulate gains. A second reason why the analyst can do better than the agent is that in many cases the analyst can have a wider perspective. Consider the

[110] This part of the discussion heavily relies on Le Grand and New (2015), Chapter 9: The Politics of Paternlism.

motorcyclist example again. Let us analyze it from the reasoning failure of limited imagination this time. It is possible that the motorcyclist cannot imagine how his life would look like should he make a terrible accident that would damage his brain for the rest of his life. Similarly, in the retirement plan example, it can be claimed that it is hard for a young person to imagine how it would be like to retire with very little or no funds at all. The reason is that in both cases many people never experienced a severe accident or retired with insufficient funds. Although the analyst himself may not have experienced these things as well, it is possible to argue that he would *suffer less* from limited imagination regarding such experiences. The reason is that in many cases the analyst observes and deals on a daily basis with people who did undergo these kinds of experiences (think for example about a government social worker that takes care of elderly people with insufficient retirement funds. If we consider a government ministry as the analyst in this case, then it arguably seems to have enough observations, enough time of listening to stories and seeing the difficulties these people are going through, that it has a good idea of how life looks like in case of under-saving for retirement. Or at least a better idea than an agent who did not have these experiences). Third, in many cases it seems that the analyst can be more objective than the individual. This relates to the reasoning failure of limited objectivity. The reason is that it is easier for a person that is not constrained by the policy to make less biased decisions. For example, a doctor is arguably in a better place to make less emotional decisions about a treatment than his patient. The simple aspect of not making a decision that concerns his own body makes him more placid and able to think clearly about the treatment effects. Consequently, he is better able to make a more objective decision. Lastly, consider the reasoning failure of limited willpower. In many cases, a third party is more able to resist momentary temptations than the individual so constrained by the policy. The reason for this is similar to the third reason: it is easier to resist emotions or temptations when it is not our own lives or bodies that are at stake. For example, if the agent would face the decision of going on a luxurious vacation now or open a savings account for his retirement with the money, he might be tempted to opt for the first (while regretting it later). Yet, for an analyst it can be easier to resist this temptation and opt for opening the savings account for retirement.

Thus, it seems that there are good reasons why the analyst can make better decisions than the individual in many cases. To be sure, I say ‘better’, in the sense that the

analyst's decision would have an overall more positive affect on the individual's well-being than without the intervention.

However, it might be objected that, still, in principle, one cannot conclude from these four reasons that it will always be the case that the analyst would do better. Perhaps. But this possibility is not a knockdown objection against BP's use of purified preferences. Even if the agent would be better off in some cases, all things considered, without the intervention, BP can avoid the second critique by implementing sufficient safeguards. For instance, democratic decision-making tools, such as public debates on policies or laws or even the use of referendums, can be helpful here. Other options are the use of a 'sunset clause' clauses. The idea of a sunset clause is that a certain law would cease after a certain period of time, unless further legislative action is taken. This can be even used together with a public debate. For example, one can install a BP policy with a sunset clause, which states that after a few years, the policy would only continue if the public votes for it in a referendum. Importantly, sunset clauses can counter some reasoning failures. For example, having a few years of experimenting with the policy neutralizes issues such as limited imagination since people get to experience how their life has changed due to the intervention. Using a sunset clause also neutralizes limited objectivity and self-control since people get ample time to reflect on the benefits or losses to their well-being due to the policy and they do not need to reply to it quickly. This is not an exclusive list, of course, of possible safeguards. For the purpose of this argument, however, it is not needed to make a complete list of such safeguards. Instead, the main claim I want to draw is that the second critique explicated in Section 2 can be sufficiently challenged by behavioral paternalists. By itself, this critique does not undermine BP's ability to use purified preferences to increase people's subjective well-being. That is, to equate purified preferences with people's own goals.

Conclusion

Behavioral paternalism equates the satisfaction of purified preferences with well-being. Thus, by satisfying purified preferences behavioral paternalists claim to increase people's well-being. Yet, can the satisfaction of purified preferences be justified as increasing people's subjective well-being? As satisfying their own goals? There are at least two reasons that raise skepticism in that regard. First, both our experience (at least

in some situations) and psychological theories do not seem to support the idea that purified preferences exist. Second, even if purified preference exists, behavioral paternalists need a lot of information to consider all the possible trade-offs such that the agent's overall well-being would not decrease.

I have argued that these critiques should not challenge the defensibility of purified preferences. In reply to the first critique, I argued that BP does not have to claim that purified preferences exist. Instead, behavioral paternalists can (and seem to) think about purified preferences as counterfactual-ifs; as the preferences that would be observed if people were fully rational. Responding in this way does not mean that BP needs to reject its means-paternalistic aspect. People can still appreciate rationality without having a purified preference that exists in their minds, and being rational in this way can have the potential to make them better off, by their own judgment. In reply to the second critique, I first argued that, in principle, since the use of purified preferences implies consistency with other goals of the agent, then satisfying purified preferences should not lead to a decrease in the agent's overall well-being. Yet, presumably, because the analyst would not always know all the agent's goals and reasoning failures, this reply might be problematic. Thus, I provided a second reply for the second critique. I argued that the analyst needs to do better than the agent, and, at least, most of the time. I then reviewed four reasons why the analyst would probably do better in many choice situations than the agent. In cases where the analyst would not, or that we suspect he would not, using sufficient safeguards can counter the remaining concerns.

Discussion

Can behavioral paternalists interfere with people's choices to make them choose one thing over another, while still holding on to the claim that by doing so people will choose the alternative that made them better off by their own subjective standard? This is the question we started with. In this last part, I shall summarize the arguments and suggest that they provide a foundation in answering positively with regard to this question.

Behavioral economics has challenged standard welfare economics in two ways. First, it showed that people are, in many situations, irrational; that in their behavior they violate axioms of rationality. I called this the descriptive challenge. But the descriptive challenge also impugns the use of preference-satisfaction as a welfare criterion. I called this the normative challenge. Based on this, we distinguished behavioral paternalism as an approach that accepts the descriptive challenge and rejects the normative one. I also claimed that there are two salient aspects that characterize BP: its use of the satisfaction of purified preferences as a welfare criterion, and it being paternalistic.

Chapter two then claimed that definitions for paternalism in economics are unsatisfactory in describing behavioral paternalism. I then suggested a new definition for BP's paternalism. In brief, a policy is behavioral paternalists if it is a) interferes with choices; b) justified by reasoning failures; c) to make the agent so constrained by it better off; and d) without infringing upon his liberty, narrowly understood. I also evaluated to what cases this definition can be extended. We saw that there are three cases on which BP purely scores: means-, broad-, and welfare-paternalism. However, it is the means-paternalistic aspect that is most important for this discussion. It reveals that behavioral paternalists are interested in interfering only with people's means and not goals.

Thus, BP wants to satisfy purified preferences and be means-paternalistic. The third chapter claimed that these two aspects reveal that BP equates purified preferences with people's goals. However, some have challenged the use of purified preferences as a valid welfare criterion that can increase people's subjective well-being. In particular, we reviewed two main challenges for the use of purified preferences. First, it is argued

that behavioral paternalists seem to think that purified preferences really exist in people's minds. But such a claim is not supported by our experiences and psychological theories. Second, using purified preferences requires a lot of information, something which is often not available to the analyst. If these critiques are convincing, then using the satisfaction of purified preferences to increase people's subjective well-being seems problematic.

I have suggested two replies to these critiques. First, behavioral paternalists need not claim that purified preferences exist in people's minds. Rather, behavioral paternalists can think about purified preferences as a tool that helps them to explain what preferences would be observed had people been fully rational. Indeed, at least some of them seem to think in this way. Importantly, thinking in this way does not necessarily undermine BP's means-paternalistic aspect. The reason is that people can (and seems to) appreciate rationality without having purified preferences that exist in their minds. Thus, they can admit that they are better off after the intervention. Second, behavioral paternalists can do better in many cases than the agent. I have presented four reasons why this may be so. In brief, in many cases, reasoning failures can be countered if the analyst has enough training or observations of similar cases. In other cases, the mere fact that the analyst is interfering with a choice that does not has to do with his own body or his life can counter reasoning failures such as limited self-control or limited objectivity. I also claimed that the use of safeguards could ensure that the overall well-being of the agent would not be decreased in cases where analysts are not able to do better than the agent.

Behavioral paternalism grabbed a lot of attention in recent discussions, both in academia and in public and private debates. To many, the use of BP policies seems like a helpful tool, as a way to improve people's well-being, judged by their own standard. Can it be that economists, who have criticized paternalism for so long, have now become some of its greatest supporters? It seems so. The arguments in this thesis suggest that it is possible to interfere with people's choices to increase their subjective well-being. Granted, it will not always be perfect. But it is not impossible as well.

Bibliography

Ainslie, George. "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control." *Psychological Bulletin* 82, no. 4 (1975): 463–96. https://doi.org/10.1037/h0076860.

Berg, Nathan, and Gerd Gigerenzer. "AS-IF BEHAVIORAL ECONOMICS: NEOCLASSICAL ECONOMICS IN DISGUISE?" *HISTORY OF ECONOMIC IDEAS* 18, no. 1 (2010): 133–65. https://doi.org/10.2307/2143802.

Bernheim, B. Douglas. "Do Households Appreciate Their Financial Vulnerabilities? An Analysis of Actions, Perceptions, and Public Policy." In *AM. COUNCIL FOR CAPITAL FORMATION, TAX POLICY FOR ECONOMIC GROWTH IN THE 1990*, 1–30, 1994.

———. "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics." *Journal of Benefit-Cost Analysis* 7, no. 1 (2016): 12–68. https://doi.org/10.1017/bca.2016.5.

Bernheim, B. Douglas, and Antonio Rangle. "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics." *The Quarterly Journal of Economics* 124, no. 1 (2009): 51–104. https://doi.org/10.1017/CBO9781107415324.004.

———. "Choice-Theoretic Foundations for Behavioral Welfare Economics." *American Economic Review* 97, no. 2 (2011): 464–70. https://doi.org/10.1093/acprof:oso/9780195328318.003.0007.

Beshears, John, James J. Choi, David Laibson, and Brigitte C. Madrian. "How Are Preferences Revealed?" *Journal of Public Economics* 92, no. 8–9 (2008): 1787–94. https://doi.org/10.1016/j.jpubeco.2008.04.010.

Blaug, Mark, and Roger Backhouse. "A History of Modern Economic Analysis." *The Economic Journal* 96, no. 382 (1986): 571. https://doi.org/10.2307/2233152.

Bleichrodt, Han, Jose Luis Pinto, and Peter P. Wakker. "Making Descriptive Use of Prospect Theory to Improve the Prescriptive Use of Expected Utility." *Management Science* 47 (2001): 1498–1514. https://doi.org/10.1287/mnsc.47.11.1498.10248.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. "Salience and Consumer Choice." *Journal of Political Economy*, 2013. https://doi.org/10.1086/673885.

Bradley, Richard. *Decision Theory with a Human Face*. Cambridge University Press, 2017.

Camerer, Colin, Samuel Issacharoff, George Loewenstein, Ted O'Donoghue, and Matthew Rabin. "Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism.'" In *University of Pennsylvania Law Review*, 2003. https://doi.org/10.2307/3312889.

Dasgupta, Partha. "Positive Freedom, Markets and the Welfare State." *Oxford Review of Economic Policy* 2, no. 2 (1986): 25–36. https://doi.org/10.1093/oxrep/2.2.25.

Dold, Malte F. "Back to Buchanan? Explorations of Welfare and Subjectivism in Behavioral Economics." *Journal of Economic Methodology* 25, no. 2 (2018): 160–78. https://doi.org/10.1080/1350178X.2017.1421770.

Dworkin, Gerald. "Moral Paternalism." *Law and Philosophy* 24, no. 3 (2005): 305–19.

———. "Paternalism: Some Second Thoughts." In *The Theory and Practice of Autonomy*, 121–29. Cambridge University Press, 1988.

———. "Paternalism." *The Monist* 56, no. 1 (1972): 64–84.

———. "Paternalism." In *The Stanford Encyclopedia of Philosophy*, 2019. https://doi.org/10.5860/choice.41sup-0181.

Evans, Jonathan St B.T., and Keith E. Stanovich. "Dual-Process Theories of Higher Cognition: Advancing the Debate." *Perspectives on Psychological Science*, 2013. https://doi.org/10.1177/1745691612460685.

Farkas, Steve, and Miles Johnson. "MILES TO GO: A STATUS REPORT ON AMERICANS' PLANS FOR RETIREMENT 9," 1997.

Feinberg, Joel. *The Moral Limits of the Criminal Law Volume 3 : Harm to Self*. Vol. 3, 1989. https://doi.org/10.1093/0195059239.001.0001.

Fletcher, Guy. *The Philosophy of Well-Being: An Introduction*. Routledge, 2016.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue. "Time Discounting and Time Preference: A Critical Review." *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice* 40, no. 2 (2002): 13–86. https://doi.org/10.1515/9781400829118-009.

Fumagalli, Roberto. "Decision Sciences and the New Case for Paternalism: Three Welfare-Related Justificatory Challenges." *Social Choice and Welfare* 47, no. 2 (2016): 459–80. https://doi.org/10.1007/s00355-016-0972-1.

Garren, David J. "Recent Work." *Philosophical Books* 43, no. 1 (2002): 5–22. https://doi.org/10.1111/1468-0149.00244.

Gert, Bernard, and Charles M. Culver. "Paternalistic Behavior." *Philosophy & Public*

Affairs 6, no. 1 (1976): 45–57.

Gert, Bernard, and Charles M Culver. "The Justification of Paternalism." *Ethics* 89, no. 2 (1979): 199–210.

Grand, Julian Le, and Bill New. *Government Paternalism: Nanny State or Helpful Friend?* Princeton and Oxford: Princeton University Press, 2015.

———. *Government Paternalism: Nanny State or Helpful Friend?* Princeton University Press, 2015.

Gul, Faruk, and Wolfgang Pesendorfer. "The Case for Mindless Economics." In *The Foundations of Positive and Normative Economics: A Handbook*, 3–42. Oxford University Press., 2008.

Hall, C. Michael. "Framing Behavioural Approaches to Understanding and Governing Sustainable Tourism Consumption: Beyond Neoliberalism, 'Nudging' and 'Green Growth'?" *Journal of Sustainable Tourism*, 2013. https://doi.org/10.1080/09669582.2013.815764.

Hausman, Daniel M. "Efficacious and Ethical Public Paternalism." *Review of Behavioral Economics* 5, no. 3–4 (2018): 261–80. https://doi.org/10.1561/105.00000090.

———. *Preference, Value, Choice, and Welfare*. Cambridge University Press, 2012.

Hausman, Daniel M., and Brynn Welch. "Debate: To Nudge or Not to Nudge." *Journal of Political Philosophy* 18, no. 1 (2010): 123–36. https://doi.org/10.1111/j.1467-9760.2009.00351.x.

Haybron, Daniel M., and Anna Alexandrova. "Paternalism in Economics." *Paternalism: Theory and Practice*, 2011, 157–77. https://doi.org/10.1017/CBO9781139179003.009.

Hayek, Friedrich A. "The Use of Knowledge in Society F . A . Hayek." *The American Economic Review*, 1945.

Infante, Gerardo, Guilhem Lecouteux, and Robert Sugden. "Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioural Welfare Economics." *Journal of Economic Methodology* 23, no. 1 (2016): 1–25. https://doi.org/10.1080/1350178X.2015.1070527.

Kahneman, D., P. P. Wakker, and R. Sarin. "Back to Bentham? Explorations of Experienced Utility." *The Quarterly Journal of Economics* 112, no. 2 (1997): 375–406. https://doi.org/10.1162/003355397555235.

Kahneman, Daniel. "A Perspective on Judgment and Choice: Mapping Bounded

Rationality. American Psychologist, 58(9), 697-720." *American Psychologist* 58, no. 9 (2003): 697–720. https://doi.org/10.1055/s-2003-44981.

———. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. "The Endowment Effect, Loss Aversion, and Status Quo Bias." *Journal of Economic Perspectives* 5, no. 1 (1991): 193–206. https://doi.org/10.1257/jep.5.1.193.

Kliebenstein, James B, Jason F Shogren, Seung Y Shin, and Dermot J Hayes. "Resolving Differences in Willingness to Pay and Willingness to Accept." *The American Economic Review* 84, no. 1 (2016): 255–70.

Kuhfuss, Laure, Raphaële Préget, Sophie Thoyer, Nick Hanley, Philippe Le Coent, and Mathieu Désolé. "Nudges, Social Norms, and Permanence in Agri-Environmental Schemes." *Land Economics* 92, no. 4 (2016): 641–55. https://doi.org/10.3368/le.92.4.641.

Loewenstein, George. "Emotions in Economic Theory and Economic Behavior." *American Economic Review* 90, no. 2 (1999): 256–60. https://doi.org/10.1257/aer.99.2.594.

Loewenstein, George, and Emily Haisley. "The Economist as Therapist: Methodological Ramifications of 'Light' Paternalism." *The Foundations of Positive and Normative Economics: A Hand Book* 1 (2007): 1–50. https://doi.org/10.1093/acprof:oso/9780195328318.003.0009.

Loomes, Graham, Chris Starmer, and Robert Sugden. "Do Anomalies Disappear in Repeated Markets." *Economic Journal* 113, no. 486 (2003): 153–66. https://doi.org/10.1111/1468-0297.00108.

McQuillin, Ben, and Robert Sugden. "Reconciling Normative and Behavioural Economics: The Problems to Be Solved." *Social Choice and Welfare* 38, no. 4 (2012): 553–67. https://doi.org/10.1007/s00355-011-0627-1.

Mill, John Stuart. *On Liberty*. London: John W. Parker and Son, West Strand., 1859.

Parfit, Derek. "What Makes Someone's Life Go Best." In *Reasons and Persons*, 1–12, 1986. https://doi.org/10.1093/019824908X.001.0001.

Pennington, Mark. "Paternalism, Behavioural Economics, Irrationality and State Failure." *European Journal of Political Theory* 18, no. 4 (2019): 565–77. https://doi.org/10.1177/1474885116647853.

Plott, Charles R. "Rational Individual Behavouir in Markets and Social Choice Processes: The Discovered Preference Hypothesis." In *Rational Foundations of*

Economic Behaviour, 1996.

Pope, Thaddeus Mason. "Counting the Dragon's Teeth and Claws: The Definition of Hard Paternalism." *Georgia State University Law Review* 20, no. 3 (2004): 659–722.

Read, Daniel, and Barbara Van Leeuwen. "Predicting Hunger: The Effects of Appetite and Delay on Choice." *Organizational Behavior and Human Decision Processes* 76, no. 2 (1998): 189–205. https://doi.org/10.1006/obhd.1998.2803.

Redelmeieer, Donald A, Paul Rozin, and Daniel Kahneman. "Cognitive and Patients' Decisions." *Jama* 270, no. 1 (2012): 72–76.

Rizzo, Mario J., and Douglas Glen Whitman. "The Knowledge Problem of New Paternalism." *Brigham Young University Law Review*, no. 4 (2009): 905–68. https://doi.org/10.2139/ssrn.1310732.

Salant, Yuval, and Ariel Rubinstein. "(A, f): Choice with Frames." *Review of Economic Studies*, 2008. https://doi.org/10.1111/j.1467-937X.2008.00510.x.

Schnellenbach, Jan. "Evolving Hierarchical Preferences and Behavioral Economic Policies." *Public Choice* 178, no. 1–2 (2019): 31–52. https://doi.org/10.1007/s11127-018-0607-4.

Sugden, Robert. "On Nudging: A Review of Nudge: Improving Decisions about Health, Wealth and Happiness by Richard H. Thaler and Cass R. Sunstein." *International Journal of the Economics of Business* 16, no. 3 (2009): 365–73. https://doi.org/10.1080/13571510903227064.

———. *The Community of Advantage: A Behavioural Economist's Defence of the Market*, 2018. https://doi.org/10.1093/oso/9780198825142.001.0001.

———. "Why Incoherent Preferences Do Not Justify Paternalism." *Constitutional Political Economy* 19, no. 3 (2008): 226–48. https://doi.org/10.1007/s10602-008-9043-7.

Sunstein, Cass R. "'Better off, as Judged by Themselves': A Comment on Evaluating Nudges." *International Review of Economics*, 2018. https://doi.org/10.1007/s12232-017-0280-9.

Sunstein, Cass R., and Richard H. Thaler. "Libertarian Paternalism Is Not an Oxymoron." *University of Chicago Law Review*, 2003. https://doi.org/10.2307/1600573.

Thaler, Richard H. *Misbehaving: The Making of Behavioral Economics*. New York: W.W Norton & Company, Inc., 2015.

Thaler, Richard H., and Cass R. Sunstein. “Libertarian Paternalism.” *American Economic Review* 93, no. 2 (2003): 175–79. https://doi.org/10.1257/000282803321947001.

———. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. *Nudge: Improving Decisions about Health, Wealth, and Happiness*, 2008. https://doi.org/10.1016/s1477-3880(15)30073-6.

Thoma, Johanna. “Merely Means Paternalist ? Prospect Theory and ‘ Debiased ’ Welfare Analysis,” 2019.

Voorhoeve, Alex. “A Response to Rabin.” In *Behavioural Public Policy*, 2012.

Wason, P. C., and J. ST B.T. Evans. “Dual Processes in Reasoning?” *Cognition* 3, no. 2 (1974): 141–54. https://doi.org/10.1016/0010-0277(74)90017-1.

Whitman, Douglas Glen, and Mario J. Rizzo. “The Problematic Welfare Standards of Behavioral Paternalism.” *Review of Philosophy and Psychology* 6, no. 3 (2015): 409–25. https://doi.org/10.1007/s13164-015-0244-5.