Review of Behavioral Economics, 2021, 8: 285–303

Boosts: A Remedy for Rizzo and Whitman's Panglossian Fatalism

Till Grüne-Yanoff

KTH Royal Institute of Technology Stockholm, Teknikringen 76, 10044 Stockholm, Sweden; gryne@kth.se

ABSTRACT

I identify two problematic conclusions that remain somewhat implicit in Rizzo and Whitman's book: the *Panglossian conclusion* that whatever the individual thinks or wants is best for her, and the *Fatalistic conclusion* that there are no justified paternalistic interventions. Against the first conclusion, I critically discuss the authors' arguments against consistency-based rationality. Against the second, I show that there is a whole class of paternalistic interventions, *Boosts*, that do not require Rizzo and Whitman's demanding epistemic preconditions in order to be successful.

Keywords: Behavioral policy, nudge, boost, Libertarianism, paternalism, rationality, consistency, knowledge deficit

JEL Codes: D90, I30

1 Introduction

Rizzo and Whitman (from hereon: R&W) are no friends of behavioral paternalist interventions. While their book offers an admirable effort to collect and systematize the many criticisms against such policies, they exaggerate tremendously in their conclusions. I will focus on two such exaggerations here. First, they argue in effect that there is no conceptual basis for ascribing welfare-relevant errors to individuals. This leads them to the *Panglossian conclusion* that whatever the individual thinks or wants is best for her. Second, they conclude that even if error-ascription were possible, paternalistic policymakers never have sufficient knowledge to justify interventions aimed at correcting these mistakes. This leads to the *Fatalistic conclusion* that paternalistic interventions are practically never justified.

ISSN 2326-6198; DOI 10.1561/105.00000143

In this paper, I argue against both these conclusions. Against the Panglossian conclusion, I show that there is a conceptual basis for ascribing error to individuals, and that a properly adjusted form of consistency-based rationality is an important component in this. Against the Fatalist conclusion, I show that there is a whole class of paternalistic interventions, *Boosts*, that can succeed without the strong epistemic requirements that R&W criticize.

2 Inclusive Rationality and the Panglossian Conclusion

R&W start their "gauntlet of challenges" with an attack on the normative foundation of behavioral paternalism. In particular, they critique the normative relevance of its underlying "puppet rationality" concept. However, the *inclusive rationality* concept that they propose instead does not provide the conceptual basis for ascribing error to individuals. Their critique thus leads them to the Panglossian conclusion that whatever the individual does must be best for them.

The predominant goal of R&W's foundational criticism of chapter 3 is to have "dispensed with the notion that mere inconsistency . . . is per se irrational" (Rizzo and Whitman, 2020, 91; when referring to their book I will from now on merely provide page numbers), and to have replaced inconsistency with an alternative notion of inclusive rationality, which puts even greater emphasis on subjective nature of normative rationality.

> In the rush to characterize certain "anomalies of choice" as violations of rationality, behavioral paternalists have been insufficiently subjectivist. Despite their stated desire to justify policy on the basis of people's own values and preferences, behavioral paternalists have unintentionally applied an external set of values – specifically, those captured by the too-restrictive neoclassical definition of rationality. Furthermore, an important part of rationality is experimenting with different choices, discovering one's preferences over time, learning from one's mistakes, structuring one's environment, adopting strategies for self-control, and working with groups of other decision-makers. These behaviors do not fit nicely into the straitjacket of "puppet" rationality, but they are perfectly sensible behaviors for real people. (17)

Whatever their specific misgivings about "puppet rationality", their core complaint always comes back to the imposition of consistency principles on preferences and beliefs. This is most obvious in their discussion of transitivity and completeness (45) and truth-tracking (121) requirements, but derived from them also criteria like framing invariance, independence of irrelevant

alternatives (80) and exponential discounting (104). R&W argue that behavioral paternalists like Sunstein and Thaler rely on the normative strength of such consistency requirements (49), which are either identical, or at least similar (52), to the neoclassical rationality axioms. These principles are part of a whole modelling tradition in economics, which might have – R&W are happy to concede – predictive or explanatory purposes (38). But they insist that for the normative purpose of determining which preferences or beliefs are welfare-relevant, this notion of rationality is useless.[1] Importantly, they do not fuss over the specifics of some neoclassical axiom; rather, they criticize all notions that conceive of rationality as the consistency of subjective attitudes.[2]

But what is this alternative rationality concept that is supposed to replace the consistency-based account? Inclusive rationality, like puppet rationality, concerns purposeful behavior based on subjective preferences and beliefs, but puts more emphasis on environmental and cognitive constraints. R&W agree that people's good is determined by the satisfaction of their actual goals, represented by their genuine beliefs and preferences (38); they also agree that only (in the correct sense) rational preferences and beliefs count towards welfare (17). Notably, and in agreement with the rejection of consistency principles, it "does not dictate the normative structure of preferences and beliefs a priori. Instead, it allows a wide range of possibilities in terms of how real people select their goals, form and revise their beliefs, structure their decisions, and conceptualize the world" (26). In particular, (i) people are free to select their goals unconstrained by their other goals, preferences or beliefs; (ii) they can change these goals at any time; (iii) they might not even have well-defined and well-articulated objectives that exist independently of choices themselves (26, 42–52, 58, 433); and (iv) their beliefs might not be truth-tracking (121).[3]

My worry about R&W's rationality concept is how it satisfies one of its normative functions: to help identify error in one's own and others' reasoning

[1] Although mainstream and behavioral economists share these rationality concepts, it is behavioral economists, to the extent that they propose paternalistic interventions, who make use of it for the purpose of determining welfare-relevance (Grüne-Yanoff, 2020b). Thus R&W critique of "puppet rationality" specifically attacks behavioral paternalists, not mainstream economists.

[2] Just to cite some passages from their book supporting this general rejection of consistency-based rationality: "inclusive rationality does not dictate the normative structure of preferences and beliefs a priori . . . Their preferences and beliefs may be inchoate, incomplete, inconsistent, mutable, and dependent on context" (26); "we have dispensed with the notion that mere inconsistency . . . is per se irrational" (91); "In the early chapters of this book, we challenged the idea that such inconsistencies *necessarily* indicate irrationality" (239).

[3] Note that consistency-based rationality does not make any substantial prescriptions or prohibitions either. Rather, it regulates the coherence of collections of subjective attitudes. In this sense, consistency-based accounts are already respecting the subjectivity of epistemic and doxastic attitudes. But R&W consider this insufficient, seeking to free rationality from these "external" constraints.

and deliberation. R&W clearly want rationality to have this function, at least for individuals learning from their own mistakes (17). But *how* inclusive rationality is supposed to perform this function, the authors say very little about, beyond stressing the importance of the successful attainment of goals:

> In a framework of inclusive rationality, the ultimate standard by which individuals' behavior is evaluated is the degree of successful attainment of goals in the actual environment in which they find themselves. (38)

This focus on actual attainment and actual consequences as the success criterion is problematic, as it disregards one of the fundamental insights of modern decision theory – the distinction between rational decision under uncertainty and the material outcome of such decisions. When deciding between different actions, each action might lead to more than one possible outcome. The relation between action and any particular outcome is typically out of the decision-maker's control – they must treat this relation as uncertain. To judge the rationality of the decision on the basis of the outcome that actually obtains is implausible, because it commends the decision-maker for a beneficial resolution of uncertainty and blames them for a disadvantageous one. But that's just good or bad luck, and therefore shouldn't be counted toward the rationality of the decision.

Imagine for example Jill and John, two individuals with the same preferences and beliefs. If Jill and John both buy lottery tickets with a small chance of winning, then Jill's decision is not more rational than John's just because her number was drawn but his wasn't. Nor is Jill more rational than John when they both ignore today's avalanche warning, and John gets himself killed but Jill doesn't. In both cases, Jill and John took the same decision, which might be rational or irrational, but the conclusions are independent of the actual outcomes that obtained. Consequently, inclusive rationality mixes rationality considerations with matters of luck, making the rationality assessment *ambiguous*.

Furthermore, R&W claim that it is the subject who determines whether a goal is successfully attained:

> A pragmatic conception of rationality demands that the suspected "irrational" behavior be shown to have consequences deemed undesirable *by the agents themselves*. (104–105, their emphasis)

But what does such a subjective anchoring mean if consistency constraints on them are not permitted? As we saw, R&W reject preference transitivity and completeness, truth-tracking requirements and beliefs and exponential discounting, to name but a few. If goals are unconstrained by other goals an agent might have, if they can freely vary in time, or if goals don't exist

independently of choices themselves, then there is no point from which a decision could be judged as unsuccessful. R&W commit themselves to such views, as when they for example, approvingly cite Buchanan's claim of the choice-dependence of preferences, concluding

> If preferences do not exist independently of the act of choice, then there is no preference set against which to judge the individual's choices as deficient. (58)

By their own argument, thus, their proposed rationality concept offers only an *arbitrary* normative standard. Judging the behavioral paternalists to have been insufficiently subjectivist, they increase the subjectivism to such an extent that conceptually anything an agent deems good or desirable must count as welfare-enhancing. This extreme subjectivism is nicely illustrated with their later discussion of the rationality of wishful thinking (bashfully described as "optimistic expectations" here):

> There is anticipatory utility associated with optimistic expectations. ... However, well-being is not maximized by ignoring the behavioral distortions (savings, investment) that will occur if expectations are not rational in the technical sense of consistency with objective probabilities. There are costs to fooling oneself. There is a trade-off between gains in anticipatory utility and the costs of behavioral distortion. (122)

In effect, here a subjective magnitude (anticipatory utility) is balanced against material consequences, and the trade-off between these two is left to another subjective judgment. Normatively, there simply is no there there – one conceptually will never be able to ascribe error to such individuals. Consequently, although the authors claim that "In principle, it is possible to make serious and systematic errors in seeking one's goals" (433), their inclusive rationality framework does not offer the conceptual tools for identifying and delineating such errors.

Adopting the inclusive rationality concept would *conceptually* prevent the identification of error in people's decision-making and force the conclusion that whatever agents deem good or desirable is best for them. This conclusion does not just arise from epistemic and practical considerations about the limitations of the policymaker – no, it arises straight from R&W's conceptualization of rationality itself. Integrative rationality implies the Panglossian conclusion.

3 Resisting the Panglossian Conclusion

One must of course entertain the possibility that Dr. Pangloss might be correct. Perhaps modern decision theory from Pascal and Bernoulli onward has been

wrong. Perhaps one cannot make welfare-relevant rationality judgments about people's decisions based on consistency principles. Or so R&W argue. Their main effort, after all, aims not so much at supporting inclusive rationality, but at rejecting "puppet rationality" – a set of consistency principles that R&W claim are used by behavioral paternalists as the foundation of their normative claims. In this section, I argue that such a rejection is not justified.

Generally, their argumentative strategy is to show that violation of these consistency principles does not *necessarily* mean the individual is irrational in the sense relevant for welfare judgments. Each successful instance of such an argument then helps reject the general welfare-relevance claim of puppet rationality. While I often find myself in agreement with R&W regarding these individual cases, I again disagree with their conclusion. Rather than generally rejecting all consistency-based accounts of (welfare-relevant) rationality on their basis, I conclude the inadequacy and oversimplication of the particular consistency principle that R&W critique, while finding other, closely related principles intuitively adequate. I will discuss a few of these instances below.

Before going into details, let me sketch my general intuition regarding why consistency considerations are welfare-relevant. An agent's welfare is determined by the degree to which she can shape the world according to her valuations (preferences, desires, values ...). Her ability of shaping the world is limited by many things like budgets, competences, and brute luck. Yet these are not the only obstacles to succeeding in shaping the world according to one's valuations. Instead, the agent's valuations and beliefs must also satisfy certain conditions: beliefs must reflect the information available about these external constraints, and valuations must provide some form of all-things-considered ranking (at least a preorder or quasi-order) of actions. These are welfare-relevant rationality conditions for the following reasons: if they are not satisfied, an agent might still obtain desirable outcomes through her actions – there are after all lucky but irrational agents, and unlucky but rational ones – but one cannot maintain that she realized these outcomes *according to* her valuations. Rationality spells out the conditions on beliefs and valuations necessary to achieve such an accord. Any account of welfare that is subjective in the sense that it takes individual goals and their influence on realizing outcomes as the ultimate criterion must therefore include such a rationality concept. These principles are thus not "external values" as R&W claim (17) – rather, they are the conditions to avoid self-defeating subjective attitudes. Therefore, some consistency-based rationality principles, whatever they might be in detail, are welfare-relevant.

My disagreement with R&W thus is not about the normative validity of particular rationality axioms in economics. I agree with them that they are often badly motivated, insufficiently sensitive to context and too simplifying. But I don't see the need to reject *all* consistency-based rationality concepts. Once one does that, as I showed in the previous section, the Panglossian

conclusion looms: the conceptual foundations for assigning decision errors disappears. But as I argue in the rest of this section, some kind of consistency-based rationality is still plausible, despite R&W's criticism. To show this, let me discuss some of their criticisms in more detail.

3.1 Preference Completeness and Transitivity

R&W reject preference completeness as a normatively relevant principle because preferences evolve through experience, trial and error:

> normatively, there is no reason to insist that, in order to be considered rational, every agent should have *already* arrived at fully consistent preferences – no more than an entrepreneur should have *already* created and implemented a full business plan, purchased all inputs, and commenced production. These are processes that play out in real time. (57)

While I agree with such an evolutionary sentiment, I am not sure what normative lesson to learn from it. To stay with the analogy, an entrepreneur at any given time might not have created a full business plan already – there is always more to explore and think of. But to give this striving any direction requires an ideal that specifies what a complete plan would be. Such an ideal of completeness is often related to considering *all relevant* eventualities (Mintzberg, 1994). This in turn allows the assessments of flawed planning: A plan that offers no guidance for likely eventualities is flawed, while a plan specifying actions for not even remotely possible eventualities is over-specified.

This equally applies to preferences. Completeness mandates the specification of preferences over *relevant* alternatives (pragmatically determined given a decision horizon) and given available information. It does not require preferences over all possible alternatives. Even if computational limitations keep people from actually having complete preferences over relevant alternatives, it remains a normative ideal, just like it is a normative ideal for entrepreneurs to develop a business plan incorporating all available information *before* they start making investment and production decisions.

Once completeness is defined over relevant options instead of all conceivable options, it is false that

> a rational person (in the inclusive sense) who compares costs and benefits will not, and *should* not, have complete and transitive preferences. (59)

Such a claim only makes sense if the set of alternatives is misspecified to include irrelevant options. If the atheist cancer patient, for example, is asked where she prefers the angels to be sitting on judgement day, she might well

insist that she doesn't care – thus it might be rational for her not have a preference regarding these options. But available palliative care options? These will most likely matter for her, affecting her comfort, lucidity, chance of recovery, and interactions with her surroundings. Achieving an outcome that – with due allowance for uncertainty – is most in accord with one's evaluations, requires consistent and complete preference over the relevant options. While most people, and perhaps in particular terminal cancer patients, might fail to satisfy these conditions, they nevertheless remain the normative ideal as long as subjective welfare matters.

R&W's criticism thus becomes moot when focusing only on preferences over *relevant* alternatives, both in terms of the number of alternatives considered, as well as in terms of the detail in which they are described. Completeness and transitivity over relevant preferences are still consistency principles, just freed from the absurd requirement of applying to *all* alternatives.

R&W also dispute transitivity as a condition over relevant options. Borrowing an example from Broome (1991), they consider Maurice, who seems to have intransitive preferences $>$ between mountaineering (M), visiting Rome (R) and staying home (H): $\mathrm{H} > \mathrm{R}$, $\mathrm{R} > \mathrm{M}$ and $\mathrm{M} > \mathrm{H}$. But then we learn that Maurice refers to different reasons when comparing these options: he finds it more relaxing to stay at home than tour Rome; he prefers cultural attractions to natural ones; and he thinks it is braver to go mountaineering than staying home. Therefore, R&W conclude, it seems entirely rational for Maurice to hold such intransitive preferences.

However, if Maurice's preferences are fueled by these considerations, then he does not have preferences over H at all, only preferences over H-when-compared-to-R (H_r) and H-when-compared-to-M (H_m). His preferences show no intransitivity at all: $\mathrm{H}_r > \mathrm{R}$, $\mathrm{R} > \mathrm{M}$ and $\mathrm{M} > \mathrm{H}_m$. *Pace* R&W, Maurice does not hold rational intransitive preferences; rather, he holds transitive (and thus rational) ones.

R&W object to this strategy of reindividuation of alternatives, again following Broome (1991), by asking "what is to stop us from redescribing the alternatives in *every* apparent case of intransitivity?" (70). It is of course true that an indiscriminate application of this strategy would make transitivity vacuous: any $\mathrm{A} > \mathrm{B} > \mathrm{C} > \mathrm{A}$ would then really be an A-when-compared-to-$\mathrm{B} > \mathrm{B} > \mathrm{C} >$ A-when-compared-to-C, and transitivity would not provide a constraint at all.

But this dismissal is too fast. A genuine subjective account should consider people's actual level of preference description. This determination is not an arbitrary redescription. At which level of description an agent compares alternatives is an observable fact (Dreier, 1996). Did Maurice actually consider the relation his choice of hiking bears on the to the alternative of staying home? If he hadn't, he would not have differentiated between H_r and H_m and his preferences would be intransitive and irrational. But if he had,

then introducing this correct level of description would prevent the mistaken impression of intransitivity. Of course, observers (and also Maurice) might have difficulties knowing what this correct level of description is. But that is an epistemic problem – conceptually, there is no difficulty.

Thus, I conclude that the plausibility considerations that R&W present against preference transitivity and completeness do not carry much weight: there *is* often something normatively inadequate about incomplete or intransitive preferences over relevant options, and this provides a *prima facie* reason to help overcome such preferences.

3.2 Intertemporal Preference Change

R&W claim that "exponential discounting is merely a modeling norm, not a prescriptive one" (104). Accordingly, an agent who deviates from exponential discounting utility (EDU) does not commit a decision error calling for correction. I agree that not every such deviation constitutes a welfare-relevant error. But I maintain that some kinds of deviations do, and that these can be at least conceptually separated from the non-welfare-relevant ones.

The key axiom underlying EDU is Stationarity. Stationarity, however, is hard to evaluate normatively – it just runs together too many disparate features to compare easily to intuition. The two necessary and jointly sufficient conditions for stationarity are *consistency* and *invariance* (Halevy, 2015, p. 342). Roughly put, consistency prohibits preference reversals due to changes in the temporal distance between agent and realization of preference. Invariance prohibits preference reversals due to absolute changes in calendar time. I submit that at least consistency under *ceteris paribus* conditions is normatively defensible. That is, preference changes effected purely by a change in delay between evaluation and relata realization, without the influence of external factors, are irrational. One has good reasons to either prevent them from occurring or from influencing one's decisions, for the following reasons (see Grüne-Yanoff, 2020a for the full argument).

One reason is that delay-dependent reversals are disempowering. People form intentions and plans, only to later feel that they do not want to realize them – without any discernable external influence (no new information, no new experience, no new insight) making it so. George Ainslie's description of addiction exemplifies such cases. There, the individual is at war with herself: conflicting intrapersonal interests, controlling behavior at different times, strategically interact to maximize their share of reward (Ainslie, 2001). The deep sense of frustration, the damage to self-estimate, the sense that one lacks the power to determine the course of one's life as a whole, I believe, is a strong *prima facie* reason to prevent delay-dependent reversals.

Another reason is that beyond the subjective experience of disempowerment, delay-dependent reversals also prevent the realization of one's plans.

Admittedly, not all long-term plans are good or desirable. But avoiding preference reversal is a necessary condition for realizing *any* long-term plan. Thus, while successful planning doesn't make projects worthwhile, projects that are worthwhile could not be realized without successful planning. The ability to realize such long-term plans is often seen as an important part of rationality (Bratman, 1983; McClennen, 1990). Avoiding preference reversal, because it is a necessary condition for the possibility of realizing long-term plans, becomes part of this rationality condition – thus making them rational in their own right.

Yet another reason is that both coordination and cooperation between persons depends on each forming beliefs about others' motivations. To coordinate actions without verbal communication requires that agents predict each other's preferences. To achieve stable cooperation (with or without communication) requires that agents assess their opponents' incentives to defect, which in turn requires them to predict each other's preferences. Intertemporal stability and the absence of preference reversals increase this predictability and thus facilitates coordination and cooperation. To the extent that coordination or cooperation is desired, avoiding preference reversals is rational.

Thus, while R&W are perhaps correct in critiquing EDU, it does not follow from their argument that there are no consistency principles on whose basis intertemporal decision error can be diagnosed. To the contrary, a necessary condition for EDU – the *c.p.*consistency principle – is a good candidate for such a normative principle. The welfare-relevance of this principle derives from the empowerment, the ability to plan and coordinate, which it facilitates. This also provides a *prima facie* reason to help people overcome violations of this principle.

3.3 Beliefs

R&W take issue with inconsistency of beliefs as a welfare-relevant rationality criterion, and in particular "challenge the idea that the sole function of beliefs is truth-tracking. . . . beliefs can provide a source of motivation to accomplish certain goals" (121), envisioning that people "may gain satisfaction purely from having a particular belief, irrespective of its truth ('My wife is beautiful and my children are gifted')" (37).

I am reading this at a time when large numbers of Americans turn away from Fox News to novice networks like Newsmax and OANN, because Fox had begun to at least occasionally question the baseless conspiracies emanating from the White House (Folkenflik, 2020). Apparently, these people prefer to believe that Trump really did win the election and move to channels that by whatever means support this belief, rather than provide the available evidence to the contrary. For R&W, these people behave rationally.

What could make such wishful thinking strategies rational? How does it support shaping the world according to one's valuations? The worldly

constraints – budget, abilities, brute luck – are of course still all there, whether one believes them or not. By making beliefs dependent on valuations, instead of on information about these factual constrains, wishful thinkers lose their guiding tool by which they can identify the actions most likely to help shape the world according to their valuations. Furthermore, because many wishful thinkers rely on others – like the above networks – to feed these 'wishful beliefs', they make themselves entirely dependent on these providers. No corrective evidence is admitted anymore, and the sole criterion becomes to what degree the beliefs match the agents' wishes. The consequence is a race towards more extreme forms of wishful thinking (as when Newsmax replaces Fox News), leading to increasing loss of guidance of one's actions by one's evaluations. These consequences, I believe are a strong *prima facie* reason, to prevent such wishful thinking beliefs.

To conclude, I argue against R&W's rejection of consistency-based rationality. Their cases, while perhaps showing that the mainstream textbook versions of rationality are too simple, do not show that consistency-based rationality is generally flawed. In particular, such consistency principles are necessary to ensure that people's valuations and beliefs don't become self-defeating: that is, losing the ability to shape the world according to them. Admittedly, to assess decisions based on such principles is often difficult *in practice* because we lack epistemic access to the features that determine relevance considerations, information availability, etc. Such a diagnostic problem, which I discuss in the next section, must be distinguished from the conceptual basis for error ascription. Conceptually, consistency-based rationality provides a clear basis for identifying welfare-relevant decision errors.

4 Overcoming the Fatalistic Conclusion through Boosts

R&W's *Fatalistic conclusion* says that even if error-ascription were possible, paternalistic policymakers never had sufficient knowledge to justify interventions aimed at correcting these mistakes. Against this, I show that there is a whole class of paternalistic interventions, *Boosts*, that do not require the high epistemic preconditions in order to be successful.

Separately from their critique of consistency-based rationality, R&W criticize paternalistic behavioral policies on epistemic grounds – namely that the policymaker rarely has the knowledge needed in order to be reasonably certain that the interference indeed is beneficial for them. Such *knowledge-deficit* arguments against paternalism consists of the following steps: First, R&W point out that the behavioral paternalists in question respect subjective evaluations as their normative basis. That is, they consider the satisfaction of a person's

genuine preferences and values to constitute what is good for that person.[4] Consequently, R&W argue, the behavioral paternalist must know the details of people's evaluations in order to justify their preferences. Yet such knowledge is unattainable in the required detail, because it "consists of subjective attitudes, perceptions, beliefs, and tastes. It includes personal strategies whose application depends on the idiosyncratic environments, routines, and social contexts of countless individuals" (279). Thus, they conclude, the paternalists do not even know what the good is that they claim to help people attain.

Second, the paternalist seeks to rectify various mistakes that people make, thus improving these people's welfare. In the previous sections I argued that there is an unambiguous conceptual basis for attributing these mistakes, but I also admitted that this notion of error often depends on subjective and contextual factors like relevance, reasons, and decision horizons. This makes it difficult to identify such mistakes in practice, and I therefore agree with R&W that policymakers often do not know whether people have committed mistakes or not.

Third, even if an error is identified, it is often unclear how an intervention can improve upon it. R&W give a convincing example (Cf. similar arguments in Sugden, 2018, p. 62):

> "If an agent shows evidence of having both Preference Set X and Preference Set Y, there is no analytical basis for designating X or Y as the 'true' underlying preference set of the agent. Maybe it's both; maybe it's neither. To choose one over the other is simply a non sequitur." (75)

In this case, intervening to promote either X or Y would require knowledge not only about the intervention itself, but also about the subject intervened on. Sometimes, knowing the subject's goals might be sufficient, but in other cases, one also needs to know how a subject would react to an intervention, and what side-effects it might produce. Like a medical doctor, behavioral policymakers would need to examine their "patients"' individual constitution and behavioral history, in order to avoid idiosyncratic side effects.

R&W argue that behavioral policymakers do not and cannot have this required knowledge. The factors that allegedly prevent people from making decisions in their own best interests are highly context-dependent, varying from person to person, place to place, and time to time; they also vary with the experimental method of how such factors are elicited. The same holds for the way people react to the various proposed interventions. Furthermore, such policies typically intervene on a population, and policymakers typically do not

[4]For example, "libertarian paternalists" seek to improve people's choices so that they are "better by their own lights" (Sunstein and Thaler, 2003, p. 1163) or what is "best as judged by themselves" (Thaler and Sunstein, 2008, p. 5).

know the distribution of the relevant properties in the population. Therefore, behavioral policymakers don't really know the ailment of their 'patients', and thus cannot really know what they are treating and what result to expect. "If behavioral paternalism is to clear the high bar its own creators have set for it – to provide an evidence-based policy program that reliably improves personal welfare from the perspective of the individual – then its practitioners need to have the vast body of knowledge we have discussed. For most proposed interventions, they do not have it." (280). As this knowledge deficit is pervasive, R&W conclude, paternalistic behavioral policies are generally not justified.

To some extent, I agree with this knowledge-deficit criticism. Consider a policy proposal like *Save More Tomorrow* (SMT; Thaler and Benartzi, 2004). SMT makes use of the fact that many people strongly dislike delaying rewards when they are close by (e.g., getting something now vs. getting double as much next week) but are more patient when such delays happen in the more distant future (getting something next year vs. getting double as much a year and 1 week from now). SMT harnesses this widespread pattern to help people save more. Imagine Larry, who refuses to increase his retirement saving when asked how much he is willing to save today (i.e., how much consumption today he is willing to give up for higher savings later). The policymaker might then redesign investment consultations in such a way that Larry instead is asked to commit *now* to making certain savings *next year*. If Larry belongs to the many people who exhibit the above-described discounting pattern, he is likely to commit to save more under the SMT program than he was willing to save before; and because inertia will likely keep him from revising this decision come next year, SMT gets Larry to save more.

SMT is widely presented as a successful exemplar of behavioral policies; yet it also raises many questions. First, why does the policymaker want Larry to save more? Unless she knows more about Larry's underlying evaluations, it is difficult to justify saving more as better for Larry in the subjective welfare framework that the behavioral policymakers are committed to. Arguably, because Larry's preferences for or against saving more are time-dependent, one might not be able to ascribe all-things evaluations over these options to him. Yet Larry might well have some higher-order evaluations – he might prefer to prefer saving more, and these 2nd order preferences might be relevant for his welfare.

Second, how does the policymaker know that Larry is making a mistake when refusing to save more? Perhaps he can expect a large inheritance, or knows that he will die young. In those cases, Larry would not make a mistake (in fact he might not be subject to the particular discounting pattern), but rather acts rationally. But if Larry's preferences for or against saving more are time-dependent, he might not have all-things evaluations over these options. That would amount to a mistake. Yet does the policymaker actually know that Larry's preferences are time-dependent?

Third, how does the policymaker know that Larry will react to the SMT intervention by saving more? Presumably, this is based on experimental observations showing that many people are sensitive to this intervention. But that evidence doesn't show that Larry, specifically, is. Instead, it might be that Larry remains unaffected, or more troubling, that he reacts in some unwanted way to the intervention – for example, he becomes suspicious of the policymaker's intention and loses trust in government institutions more generally. Furthermore, policymakers don't know whether Larry has devised his own ways to overcome preference instability, and whether their intervention might undermine such self-controlling strategies, actually making things worse for Larry. The point here, again, is that the policymaker doesn't know. And unlike a doctor who can check with the specific patient for allergies or other adverse indications before administering a treatment, the behavioral policymaker typically does not and cannot acquire this individual-specific information.

I thus agree with R&W that the implementation of many behavioral policies is problematic due to knowledge deficits of the policy maker. I part ways with them, however, in their claim that knowledge deficits are pervasive and *generally* leave paternalistic behavioral policies unjustified: "[they] need to have the vast body of knowledge we have discussed. For most proposed interventions, they do not have it." (280). This I call their *Fatalistic conclusion* – that the paternalistic policymakers almost never had sufficient knowledge to justify interventions aimed at correcting mistakes.

My first argument against the fatalistic conclusion is to point to our ability to sometimes acquire the relevant knowledge. Many social scientists in recent years have increasingly focused on developing better tools to identify cognitive mechanisms and mental attitudes. In some cases, knowledge of cognitive mechanisms a subject engages with is sufficient to diagnose an error and recommend paths for its correction (Grüne-Yanoff, 2020b). Against these developments, R&W's general dismissal of verbal statements, regret, self-commitment or planning – as "weak defenses" (77) is not very convincing.[5] Thus at least sometimes, behavioral paternalists and more specifically nudgers have the prerequisite knowledge to justify their interventions.

My second argument against the fatalistic conclusion is to point to a category of paternalistic interventions that do not make the epistemic requirements that R&W wrongly ascribe to all behavioral interventions. In fact, while knowledge deficits are an important problem to reckon with, there are many ways to design paternalistic interventions that avoid it. One might design interventions in such a way that they are *innocuous* for those not targeted. Policymakers in the SMT example do not need to worry about whether Larry

[5]Nor is their insistence on the conceptual purity of revealed preference theory for the sake of rejecting "the admixture of mental and subjective concepts with the 'principle of revealed preference'" (86–87).

made a mistake in refusing to save more – if he had a stable preference to not save more, then the SMT intervention would not convince him otherwise. Behavioral policy proponents have argued that their proposed interventions indeed are innocuous, for example in that they are "asymmetric" – they only affect those who behave irrationally – (Camerer *et al.*, 2003, 102) or in that they are "cheap to avoid" – people who don't want to be affected can opt out without any substantial costs (Thaler and Sunstein, 2008, p. 6). Unfortunately, there is little evidence that they all are innocuous, and providing evidence for such claims might itself require particular and contextual knowledge.

However, we recently proposed a categorization of behavioral policies into different kinds (Hertwig and Grüne-Yanoff, 2017), and in particular distinguished Nudges and Boosts. Boosts, I now argue, are by their nature much more innocuous than nudges, and furthermore do not require the particular, detail-rich knowledge that R&W claim policymakers rarely have. Boosts are behavioral interventions that "foster competences through changes in skills, knowledge, decision tools, or external environment" (Hertwig and Grüne-Yanoff, 2017, p. 974). A competence might consist in many different properties, but can be broadly characterized as a "roughly specialized system of abilities, proficiencies, or skills that are necessary or sufficient to reach a specific goal" (Weinert, 2003, 45). Crucially, competences characterized this way are part of the agent's deliberational process, and do not have an influence on people's behavior without it.

To illustrate, consider the fact that many people overestimate the magnitude of a risk presented in relative terms (e.g., "a 20% decrease in fatalities"). A boost intervention would train agents to always translate statistical information they encounter from a relative probability format into a natural frequency format (e.g., "a reduction of 200 fatalities in a population of 1000", Sedlmeier and Gigerenzer, 2001). The intervention consists in training this translation competence, not in favoring one presentation format over another – after all, the natural frequency format alone might also affect the agent's understanding negatively. That way, the agent can appreciate that there are different modes of representing risk, and also that these different modes have an influence on their understanding of risk, thus allowing them to form a more reflected assessment of the risk. A nudge (e.g., Malenka *et al.*, 1993), in contrast, would choose to present the information in that format which is expected to yield the desired framing effect, without necessarily teaching the decision-maker any competence.

Boosts aim to change behavior by intervening on agents' cognitive heuristics. In the first place, this requires that boost proponents identify a deficit: for example, that people apply heuristics that yield less successful behavioral results. Without such an argument, developing and implementing a boost would be unmotivated. In this, they start out similarly as nudge proponents. But such a motivating identification is *general* and *conjectural*: as a reason for a boost, it suffices to argue that some people in some situations might

make such a mistake. For the nudge proponent, this is not sufficient: she must argue (with high confidence) that *for the particular population* for which the intervention is proposed, people make this mistake. For if this were not the case, an effective nudge would change people's behavior, although there was nothing wrong with that behavior (of those people, in that context) in the first place. Boosts do not face this problem: they train people in more effective heuristics, but leave it to individual agents when to apply them. It is thus the individual's responsibility, and not the boost proponent's, to assess whether she uses a suboptimal heuristic in a particular context. Thus boosters, in contrast to nudgers, can avoid the difficult question whether particular people in a particular situation systematically commit mistakes or not.

Boosts are innocuous, because they *must* coopt subjects' motivation to be effective. Consider an illustration concerning the above risk-format translating boost. Laura participated in such a boost training session, and now faces an important decision between two drugs, both of which present their effectiveness in terms of a reduction of relative risks. Laura had to make at least three choices regarding this boost intervention: (i) whether to participate in the training session, (ii) whether to accept the skill trained in that training session, and (iii) whether to apply this skill to the particular decision between those drugs. Consequently, only those who consider themselves in need of such competence boosts will choose to listen, learn and apply. This acts as a subject-centered, context-specific stopgap that any intervention must pass to be implemented. Although it does not guarantee innocuousness – subjects, after all, might be wrong about their own needs – it makes it much more likely that boosts are indeed interventions that are innocuous to those not targeted (Grüne-Yanoff, 2018).

Finally, Boosts' coopting of motivation also avoids the knowledge-deficit problem: the coopted motivations incorporate subjects' specific, situational, contextual knowledge. Because it is left to subjects to choose between being boosted or not, this specific knowledge – which critics argue policymakers typically have access to – is provided by the subjects themselves. Boosting policymakers therefore need not worry about whether specific individuals already have certain competences, whether boosting a specific competence would interfere with her existing competences, or whether a certain competence is useless to her. They simply leave those particular worries to the involved individuals, who can draw on their privileged specific knowledge. The boosting policymaker instead only has to ensure that for the population on average, such a boost would be useful, desirable, and not undermining; and such general pattern knowledge can typically be acquired through standard social science research. Boosts thus offer paternalistic interventions that avoid the knowledge-deficit criticism, simply because they do not require the kind of specific knowledge that the critics claim is typically not accessible to behavioral policymakers.

R&W suggest that interventions like Boosts are not paternalistic (414–419). But that uses a rather restrictive notion of paternalism. Boosts are justified by the diagnosis that people often make mistakes, and that they would be better off without these mistakes. Boosts thus rely on a conceptual basis of error ascription just like nudges; and as they intervene with the aim to overcome these errors for the benefit of those committing them, they are paternalistic. Furthermore, although boosts are often very cheap, they nevertheless impose material costs: for example for experts engaged, awareness raised, training provided and results analyzed. These costs are imposed by the policymaker who judges that people are better off when boosted – thus deciding about people's benefit on their behalf. Finally, boosts might also impose immaterial costs. By making people aware that they might make mistakes, and inviting them to contemplate their own ignorance, boosts might be seen by some as meddlesome and annoying. This all makes Boosts much more than mere information provision – instead they are paternalistic interventions that overcome B&W's epistemic criticism and thus show their Fatalistic conclusion to be unfounded.

5 Conclusion

R&W's "gauntlet of challenges" to behavioral paternalism is interestingly ambiguous. Is it a chivalrous challenge to a fair debate? Or the murderous "running the gauntlet", almost inevitably ensuring the adversary's demise? I often got the impression that it was the latter, and noted various illicit weapons amongst their defilé of arguments.

One was the subjectivism leading to the *Panglossian conclusion*. By withdrawing the conceptual basis for error ascription, it becomes inevitable to conclude that whatever the individual thinks or wants is best for her, and the behavioral paternalist has nothing left to do. Against this conclusion, I argued that R&W's cases do not show that consistency-based rationality is generally flawed, and that many intuitive notions of decision errors indeed require consistency-based principles. Thus, although standard neoclassical versions of rationality might be deficient, there are viable conceptual bases for error ascription, and the Panglossian conclusion is not valid.

Another was R&W's argument that paternalistic policymakers never had sufficient knowledge to justify interventions aimed at correcting these mistakes. This led to the *Fatalistic conclusion* that there are no justified paternalistic interventions. Against this conclusion, I showed that there is a whole class of paternalistic interventions, *Boosts*, that do not require B&W's demanding epistemic preconditions in order to be successful.

What remains is a valuable collection of relevant criticisms. Paternalistic interventions deserve scrutiny, and their easily imagined abuse must be checked. But there are no grounds for ruling out in principle attempts to improve others'

decisions in their own interest, nor for throwing in the towel on the design and application of such policies on epistemic grounds. We should strive to escape *inapt* paternalism, not paternalism generally – and boosts are a promising path of doing so.

References

Ainslie, G. 2001. *Breakdown of Will*. Cambridge: Cambridge University Press.

Bratman, M. 1983. "Taking plans seriously". *Social Theory and Practice*. 9(2/3): 271–287.

Broome, J. 1991. *Weighing Goods: Equality, Uncertainty and Time*. Hoboken, NJ: John Wiley & Sons.

Dreier, J. 1996. "Rational preference: Decision theory as a theory of practical rationality". *Theory and Decision*. 40(3): 249–276.

Folkenflik, D. 2020. "Newsmax Rises On Wave Of Resentment Toward Media – Especially Fox News". *NPR*. URL: https://www.npr.org/2020/11/30/939030504/newsmax-rises-on-wave-of-resentment-toward-media-especially-fox-news.

Grüne-Yanoff, T. 2018. "Boosts vs. Nudges from a welfarist perspective". *Revue d'économie politique*. 128(2): 209–224.

Grüne-Yanoff, T. 2020a. "In defense of intertemporal consistency. A discussion of Craig callender's 'the normative standard for future discounting'". *Australasian Philosophical Review*. (in press).

Grüne-Yanoff, T. 2020b. *What Preferences for Behavioral Welfare Economics? Journal of Economic Methodology*, accepted 26.08.2021.

Halevy, Y. 2015. "Time consistency: Stationarity and time invariance". *Econometrica*. 83(1): 335–352.

Hertwig, R. and T. Grüne-Yanoff. 2017. "Nudging and boosting: Steering or empowering good decisions". *Perspectives on Psychological Science*. 12(6): 973–986.

Malenka, D. J., J. A. Baron, S. Johansen, J. W. Wahrenberger, and J. M. Ross. 1993. "The framing effect of relative and absolute risk". *Journal of General Internal Medicine*. 8(10): 543–548.

McClennen, E. 1990. *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.

Mintzberg, H. 1994. *The Rise and Fall of Strategic Planning*. New York, NY: The Free Press.

Sedlmeier, P. and G. Gigerenzer. 2001. "Teaching Bayesian reasoning in less than two hours". *Journal of Experimental Psychology: General*. 130(3): 380–400.

Sugden, R. 2018. *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford: Oxford University Press.

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

Thaler, R. and C. R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth and Happiness*. New York, NY: Simon and Schuster.

Thaler, R. H. and S. Benartzi. 2004. "Save more tomorrow™: Using behavioral economics to increase employee saving". *Journal of Political Economy*. 112(S1): S164–S187.