Review of Behavioral Economics, 2021, 8: 379–394

Inclusive Rationality and Paternalism: Responses to Comments and Criticism

Glen Whitman[1] and Mario J. Rizzo[2]

[1] *Department of Economics, California State University, Northridge, Northridge, CA, USA; glen.whitman@csun.edu*
[2] *Department of Economics, New York University, New York, NY, USA; mario.rizzo@nyu.edu*

ABSTRACT

The symposium articles fall roughly into three groups: those that connect our work to that of other thinkers and research programs; those that raise concerns about our analytical approach; and those try to delineate principles to guide behavioral policymaking. We will consider these three groups in turn.

We wish to thank all the participants to the symposium for their contributions. In particular, we thank Nick Cowen and Malte Dold for taking the initiative to organize the symposium and to edit the papers. We appreciate the complimentary responses to *Escaping Paternalism* (hereafter EP), of course, but perhaps the critical responses even more. This is the level of engagement we hoped for when writing the book. Some of the articles in the symposium would warrant a dedicated point-by-point reply, but here we will have to be more brief.

The symposium articles fall roughly into three groups: those that connect our work to that of other thinkers and research programs; those that raise concerns about our analytical approach; and those try to delineate principles to guide behavioral policymaking. We will consider these three groups in turn.

Connections to Other Thinkers and Research Programs (Matson & Dold, Peart, Koppl)

Matson and Dold (2021) find a variety of connections between EP and the thought of David Hume. The overall theme is that there exist multiple conceptions of the good life, which in economic terms will take the form of different preferences. Furthermore, the philosopher does not occupy a privileged

ISSN 2326-6198; DOI 10.1561/105.00000148

position in arbitrating amongst such conceptions, but instead necessarily brings his own biases and preconceptions to the discussion. Aside from supporting our position that behavioral economists ought to think of their findings "as friendly advice directly to their coequal fellow citizens, not as if to a benevolent autocrat" (Matson and Dold, 2021), the Humean position here also helps avoid what Koppl (2021) calls the "anthill problem" – about which more below.

We have little to add to Matson & Dold's insightful article except to say we agree. It is, though, perhaps unavoidable that social scientists will be called upon to advise policymakers. And we, too, have engaged in some policy advice. The question is how such advice should be leavened by Hume's lesson. The answer, we suggest, is to offer advice with a strong dose of humility and a proper respect for the ability of people to have different views of the good life. We must beware the tendency to read our own predispositions into the preferences and behaviors of others.

Peart (2021) finds important commonalities between EP and the perspectives of John Stuart Mill and Philip Wicksteed (both broadly anti-paternalist) and stark differences from that of William Stanley Jevons (paternalist). Although we were familiar with Mill's work before, we want to thank Peart for drawing our attention to Mill's emphasis on the heterogeneity and complexity of human thought and behavior. These are themes that do not get enough attention. The neoclassical form of rationality is, in its essence, an attempt to capture human choice in an abstract and clear-cut form: People have a scale of values; they enact the scale subject to constraints; end of story. It is essentially a black-box form of analysis. Behavioral economists have added some complexity and opened the box to some extent – but their conclusions are still often cast as additional parameters in the same (now broken) optimization problem: two rates of discount instead of one, a larger weight on losses than equivalent gains, and so on. Such abstractions can surely be useful for positive purposes. The error is to superimpose an abstract pattern on the complexity of actual thought, and then to judge the latter deficient for not fitting the pattern. Recognizing the legitimate heterogeneity and complexity of real people's thought, Mill would have resisted the abstract and axiomatic version of rationality. It appears that Wicksteed did reject it, anticipating some of our arguments in defense of intransitive preference orderings.

We were also unaware of the extent of Jevons's paternalism. Based on Peart's description, Jevons's views presage those of behavioral paternalists of today (with the added element of critical judgment aimed at the lower classes, which is more hidden today). Particularly striking is Jevons's apparently willingness to disparage less-patient behaviors as "systematic mistakes" rather than the implementation of a higher rate of subjective time preference. In this, Jevons foreshadows what we have called the "non sequitur at the heart of behavioral paternalism" (EP, 75): identifying an apparent intrapersonal

preference conflict, and then resolving that conflict in favor of one side – usually the one with greater social approval.

Koppl connects our work in EP, not to specific thinkers, but to a broader research program that he dubs "the zoological perspective." He says that our argument – with which he generally agrees – would be strengthened by integrating it with biological evolution. Such an approach would offer an explanation for at least some of the alleged "anomalies" of human behavior and cognition identified by behavioral economists, showing how they were in fact adaptive under certain conditions. Simultaneously, the zoological perspective allows us to understand the controlling impulses of some paternalists as an attempt to gain dominance and status through prestige.

We think Koppl's suggestion is insightful, and we tend to agree. We would point out, however, that some behavioral paternalists believe they have a zoological perspective already. A common argument is that humans have cognitive traits that evolved in our ancestral past as adaptations to conditions that no longer exist. For instance, we have an inborn attraction to sweets because, in the environment in which early humans evolved, sugar was relatively rare and hard to come by. Now we live in a time of cheap and abundant sugar, but our genomes haven't caught up, and so we continue to seek out "too much" sugar. We think this argument is essentially correct, except for the unnecessary element of normative judgment. Regardless of the origin of our love of sugar, the fact is that we love it now, and we have little ability to change that. Similarly, we enjoy sex because it allowed our ancestors to procreate (and, some have argued, to bond productively with other group members). In the modern era, less procreation is necessary. Nevertheless, we still enjoy sex, and we may seek it out without any desire to procreate or bond. Our desires for sugar and sex have evolutionary roots, but we still want them. These are legitimate preferences now, *even if they are not adapted to the current environment*, and it is perfectly rational to act upon them. If sugar is cheaper now, then it makes sense to consume more sugar.

We suspect Koppl would agree on this point; an evolutionary perspective provides positive insight, but not normative conclusions. Interestingly, his argument about prestige and dominance allows us to understand why some experts may nevertheless attempt to leverage knowledge about "maladapted traits" to gain status via control over others. The impulse is understandable from an evolutionary perspective, and we might even wish to indulge it, if only doing so didn't require subjugating the valid preferences and choices of others. We hasten to add that we are reluctant to attribute motives of dominance to our intellectual opponents, as it verges on *ad hominem*. Our purpose here, as Koppl suggests, should rather be to avoid the anthill problem by reminding the supposed experts that they are ants like the rest of us. As we argue at the end of EP, their behavioral insights serve us best when offered not as edicts or

manipulations from superiors, but as friendly advice to similar people facing similar problems – a point echoed by Matson & Dold as well.

Concerns about the Analytical Approach (Rajagopalan, Grüne-Yanoff)

Rajagopalan and Grüne-Yanoff raise closely related concerns about inclusive rationality, which we present in EP as an alternative to the axiomatic rationality shared by neoclassical and behavioral economics. Rajagopalan is generally sympathetic to inclusive rationality as a richer approach that allows us to better "understand the human condition, instead of conditioning humans to fit better into existing models" (2021). But she raises concerns about whether inclusive rationality may be so broad as to be unfalsifiable (2021). The concern is that, because inclusive rationality seemingly rules nothing *out*, it may be impossible in principle to find evidence that contradicts it.

Grüne-Yanoff (hereafter G-Y) does not mention falsification, but his concern is similar. He argues that inclusive rationality is Panglossian, in the sense that it implies "that whatever the individual does must be best for them" (2021). By allowing agents to have intransitive preferences, non-truth-tracking beliefs, and so on, G-Y says we have effectively ruled out any possibility of true failure. This concern mirrors Rajagopalan's, inasmuch as both are looking for the exceptions to inclusive rationality. Such exceptions would both (1) allow for falsification and (2) avoid the Panglossian conclusion.

As Rajagopalan acknowledges, we anticipated many of these concerns in EP. So if our counterarguments there are not convincing, we may have little left to offer. But we will endeavor to make our case more clearly and persuasively.

If there is one overarching lesson of EP, particularly in the first few chapters, it is that *we should not mistake positive concepts for normative ones.* In the book, we repeatedly emphasize that neoclassical rationality can and often does serve a useful function. It allows for the creation of elegant mathematical models. At least some of the time, such models provide a reasonable approximation of how people really behave. Behavioral exceptions to neoclassical rationality, such as framing or endowment effects, can serve the same function. We applaud these efforts. As a positive matter, our position is as scientific as you can get: *Make falsifiable hypotheses! Go out and test them!* But in doing so, *do not assume your hypotheses have normative import.*

We are reminded of the old joke about a man searching for his keys under the lamp post, instead of the bushes where he dropped them, because "the light is better over here." Science wants to be testable and falsifiable. That leads us to adopt working concepts that limit our view to things that can be observed, measured, and verified. But many things that can be observed, measured, and verified may lack normative significance, particularly in the social sciences. Why? Because it's much easier to observe someone's choices

than to observe their *true reasons* for doing what they do. The former is easier to test. The latter is what matters for normative judgment.

This is why behavioral economics has focused on testing (and falsifying) consistency requirements of the sort that G-Y defends. We can't usually observe whether someone truly prefers A to B or the reverse, but we can certainly observe that they choose A sometimes and B other times – which *seems* like a contradiction. And then the temptation is to apply *that* observation to the more interesting question of whether the agent has made a normative error. But that leap is simply unjustified. It would be like assuming that any key found under the lamp post must be the one that opens our front door.

G-Y spends a good deal of time defending some form of consistency as a rationality criterion. We do not have the space here to reply to all of his arguments. But in many of his arguments, what we detect is a need to bend positive analysis to normative purposes. Effectively, the claim is that *if we do not require people to be consistent, then we lose our means of judging them.* Or, perhaps, G-Y might say that *they have no means of judging themselves.*

For instance, G-Y says, "Achieving an outcome that – with due allowance for uncertainty – is most in accord with one's evaluations, requires consistent and complete preference[s] over the relevant options" (G-Y 2021). So the problem, as G-Y sees it, is that without a consistent ranking of options, we cannot say whether the individual has in fact succeeded in achieving her values. But is this a problem for the agent, or for the economist? We say it's the latter. If the individual does not have a consistent ranking of A and B, then we cannot prove she's making a mistake when she chooses A over B (or the reverse). That's very frustrating if our goal is to identify mistakes or the absence thereof. The agent is refusing to let us judge her! But we deny that it's irrational for the agent to lack such a ranking, even if it makes economists' work more difficult.

Now, it would certainly be very strange if a person had *no* ranking over anything whatsoever. That would be tantamount to lacking goals and values entirely. Sometimes this seem to be G-Y's concern – but that is not our position. Our view of human preferences is that people have "regions" with greater consistency and structure, but also "regions" that are more protean and amorphous – i.e., where they haven't decided, considered, or explored enough to have achieved a consistent ranking. It's also worth noting that some objects of choice may be ranked with respect to some alternatives but not others. To take a simple example from Whitman (2021), "even if I have an intransitive ordering of pizza, hamburgers, and spaghetti, I may nevertheless definitively rank all three above sushi" (6). When an individual lacks a definite ranking, yet realizes that a particular decision would benefit from having such a ranking, she may incur costs for the option value of deciding later, by which point a definitive ranking may have emerged (for an example see EP, 89–90). Point being, as long as a person has at least some settled preferences, that

can be enough to motivate decisions – and moreover, *some decisions can be understood as rational coping with inconsistent or indeterminate preferences.*

G-Y concedes that consistency axioms *as applied* have tended to be "badly motivated, insufficiently sensitive to context and too simplifying" (2021). However, he argues that on some deep level the individual must have complete and transitive preferences, at least over "relevant options" (2021). Set aside how much work the word "relevant" is doing in this context. G-Y's approach is what we have dubbed the "redescription" strategy. The idea is that seemingly inconsistent preferences can be made consistent by building more into the description of the choice options.

Of course, this strategy is too powerful. As G-Y admits, it could literally dissolve any inconsistency. G-Y's answer is to focus attention on whether the individual really does construe her options in the manner that would rationalize them; this, he says, is "an observable fact." What G-Y has in mind is, in essence, a survey or interview to ascertain what the individual was really thinking when he made his choices. In other words, we ask him to justify his preferences, and then we decide whether his reasons are good enough. However, if we judge that his reasons are good enough, then what is gained for normative purposes in forcing his decisions into a transitive frame? Dreier, whom G-Y cites on this point, puts it this way: "if, for example, he simply hadn't noticed that his pairwise preferences . . . were intransitive, or if he did notice it but did not care – then after all his preferences are irrational" (1996, 260). So apparently *noticing and caring about transitivity* is now the normative standard. This rules out, for instance, deciding that an intransitivity in one's preferences is not worth the effort of resolving because not enough is at stake. We find this unpersuasive.[1]

But even if we assume that such mental evaluations are *in principle* observable, usually they are not *in fact* observed. Mental states are hard to access (except for those experiencing them). Surveys and interviews are costly to perform and may not elicit truthful answers. Interpreting and processing the results objectively is challenging at best. Behavior, on the other hand, can be seen and measured. And this is precisely why so much social scientific research, including in behavioral economics, is performed by observing behavior and then – if normative conclusions are desired – measuring it against some external standard such as consistency.

This brings us back to falsifiability. The most normatively relevant facts are often the most difficult to ascertain; they lie in the poorly lit bushes. Positive research is most effective at finding things under the streetlamp. When we're dealing with a complex phenomenon, as human choice surely is, we must accept a lower degree of falsifiability precisely because not all of the

[1]For some examples of ecologically adapted rules that may, nonetheless, result in intransitive choices, see Rizzo (2019).

interrelated variables are measurable (Hayek, 1967). When we admit that the facts do not give us reliable proof of people's irrationality, that does not make us Panglossians; it makes us agnostics.

All that said, we do not intend inclusive rationality to function *exclusively* as a normative concept. We intend it as a positive research program – that is, an organizing principle for generating testable hypotheses, rather than being directly testable itself. It suggests that, for a wide range of human behaviors, we will eventually find explanations that fit the pattern of "purposive choice that serves subjective ends subject to constraints." This encompasses many possibilities, some of which may be proved wrong. We would discard the research program if it ceased to be productive – i.e., it stopped generating useful and sometimes successful hypotheses – and if an increasing number of phenomena escaped its explanatory reach, despite efforts to include them. But in writing the book, we found the opposite: if you look for a reasonable explanation, you'll very often find one.[2]

This brings us to Rajagopalan, who suggests some principles that would help guide a productive research program in the area of inclusive rationality. We broadly concur with Rajagopalan's suggestions, particularly that we need a language for discussing struggle and aspiration that avoids both rationality-by-assumption (as in the neoclassical approach) and the presumed irrationality of all inconsistency (as in the behavioral approach). The chasm between those two can only be resolved by getting inside people's heads as much as is feasible. Doing so could allow us to elucidate factors that would identify irrationality in a manner unrelated to the violation of consistency axioms.

We are particularly intrigued by the notion that irrationality might be associated with tastes or inclinations that, rather than confronting other tastes or inclinations with which they conflict, instead avoid the confrontation and issue directly in action. As Callard (quoted by Rajagopalan) says, "The akratic's intrinsic conflict prevents these two reasons from being *in conversation with one another*" (2021, emphasis added). Although we would resist characterizing every "akratic" person in this way, there may well be individuals whose conflicting impulses never talk to each other. Instead, their struggle manifests as a kind of tragedy of the commons wherein competing impulses struggle for control in a way that undermines both. In this approach, the problem is *not* inconsistency, which the individual might be fine with. The problem is a failure

[2]Compare this to the theory of evolution. Evolution is a research program. It would be difficult to find observations that falsify evolution in its broadest sense. What can be falsified are more specific hypotheses about how certain classes of organisms evolved, what survival function particular traits served, and so on. Such hypotheses are tested and falsified on a regular basis. What would falsify evolution as an overarching research program? As a complex phenomenon, it is subject to a different degree of falsification. We would reject evolution as a research program if it ceased generating fruitful explanations over a wide range of cases. See Coyne (2009, pp. 17–18) for a list of testable hypotheses generated by the evolutionary research program.

to have an internal negotiation. See Whitman (2006) for a related discussion of how internal conflict can be treated as a Coasean bargaining problem.

We are also intrigued by the notion that an individual, through a series of individually rational choices, might nevertheless find herself in a dead-end that defies escape. Addiction might fit this model. In Rajagopalan's description, "Maybe individuals initially believe it is a rational decision, but as the addiction progresses it may reduce their ability to control their consumption at a level where the benefits exceed future costs" (2021). In essence, the person has gambled with her future ability to weigh costs and benefits – and the gamble turned out badly. Alternatively, the dead-end might result from a version of hill-climbing that leads to a local but not global maximum. A local maximum cannot be escaped by further hill-climbing, only by a bold decision to travel through a suboptimal valley *en route* to a higher hilltop.

There is a natural connection here to Cowen and Trantidis's discussion of Hayek's evolutionary account of individual cognition, wherein "rational consciousness is an emergent outcome of piecemeal adaptation to the phenomena we encounter" (2021). It is worth noting that Hayek also allowed for the possibility of dead-end paths in his evolutionary account of the common law: "The development of case-law is in some respects a sort of one-way street: when it has already moved a considerable distance in one direction, it often cannot retrace its steps when some implications of earlier decisions are seen to be clearly undesirable" (1973, 88). The idea that similar path-dependent processes could afflict individual decision making seems like a fruitful avenue of research.

We should emphasize that either of these research paths might be mistaken. We endorse them because they fall broadly within the inclusive rationality research program, not because they are necessarily correct. Further, we also should not be guided exclusively by the *falsifying* goal of finding exceptions and "edge cases." On the contrary, we should also engage in the *confirming* goal of finding more varieties of inclusive rationality. Falsification, if it occurs, would take the form of our looking for inclusive rationality and repeatedly not finding it.[3]

[3]A simplistic understanding of Karl Popper's falsificationism might lead some to think confirmatory evidence is contrary to the scientific method, but this view is mistaken. Again, an analogy to the theory of evolution is helpful. One specific prediction of evolutionary theory relates to the origin of marsupials. The first marsupial fossils appear in what are now the Americas, dating 40–80 million years old. The oldest marsupial fossils in Australia date to only 30 million years ago. During the time gap, marsupials must have found a path to Australia – and during the period in question, South America was connected to Australia by way of what is now Antarctica. This led to the prediction that marsupial fossils between 30 and 40 million years old would be found in Antarctica. Sure enough, paleontologists eventually did find marsupial fossils 35–40 million years old in Antarctica (Coyne, 2009, pp. 94–95). In a case like this, testing involved an attempt at *confirmation* (looking for fossils). Falsification would have resulted from repeated failure to find confirmation despite reasonable efforts.

Delineating Principles to Guide Behavioral Policymaking (Grüne-Yanoff, Cowen & Trantidis, Hands, Hargreaves Heap)

Aside from his analytical concerns, G-Y also takes issue with our "fatalistic conclusion" that "paternalistic interventions are practically never justified" (2021). Although G-Y concurs with much of our analysis about the pitfalls of paternalistic intervention, he suggests a category of interventions that might nevertheless pass muster: *boosts*. Boosts are intended to "foster competences through changes in skills, knowledge, decision tools, or external environment" (Hertwig and Grüne-Yanoff, 2017, p. 974). Paradigmatic boosts include training people to translate relative probabilities into natural frequencies and teaching them tools for improving motivation and self-control (Hertwig and Grüne-Yanoff, 2017, p. 979).

With a few provisos, we are generally amenable to the "boost" perspective. Throughout EP, we emphasize that inclusive rationality encompasses a panoply of techniques that real people use to control their own biases and shape their own behavior, such as joining support groups, enlisting the support of family and friends, seeking advice from experts, reading self-help manuals, strategically shaping their environments, making use of market-provided tools, and relying on group decisionmaking (EP, 218–220). In many ways, boosts sound like more strategies to add to our list. It's also notable that boosts do not intentionally push a person toward particular decisions in particular cases; they simply aim to foster competence, which the individual can choose when to deploy. In that sense, boosts seem a lot like education.

But G-Y insists that boosts are indeed paternalistic, inasmuch as they ascribe error to agents and "intervene with the aim to overcome these errors for the benefit of those committing them" (2021). Obviously, there is a semantic element here. Paternalism is not perfectly delineated, and surely there are boundary cases. But we resist applying the word paternalism to every case of people merely trying to help others – or even just themselves. We find it peculiar and misleading to class as paternalism things that could as easily be described as "helpfulness," "good customer service," "ergonomic product design," or "responsiveness to consumer demand" (EP, 414). When the term paternalism is applied so loosely that it includes using a GPS, giving someone directions, having a low-calorie section on a menu, or putting reminders in your calendar, we submit that it has been stretched so far as to be meaningless. In all these activities, we see not paternalism, but the *alternatives* to paternalism. Furthermore, we believe including them under the same umbrella is part of a deliberate strategy by behavioral paternalists to elide important distinctions, thereby easing the transition to more overtly paternalistic measures (EP, 395).

But setting aside the semantics, the real question is *how* boost advocates intend to implement their proposals. The boost literature is not entirely clear on this point. Many boosts involve some kind of training – in cognitive methods,

in life skills, and so on. Would people be required to take such training? Would it be mandated by the government, or instituted by educational institutions, or provided by firms? We aren't sure, but we think the answers matter.

Fortunately, boost advocates do emphasize that boosts "require the individual's active cooperation" and that "Individuals choose to engage or not to engage with a boost" (Hertwig and Grüne-Yanoff, 2017, p. 982). This suggests that boosts would generally be offered on a voluntary basis, at least for adults. In that sense, they fall nicely within our list of ways people try to manage their own biases and behavior. In the final chapter of EP, we offer a list of key distinctions that we believe will help contain the most problematic interventions. Among those distinctions are "self-imposed versus other-imposed," "invited versus uninvited," "informative versus manipulative" (EP, 414–417). If our reading of G-Y and Hertwig is correct, then boosts would fall on the better side of these divides.

Cowen and Trantidis (hereafter C&T) largely concur with our analytical approach. Based on arguments in EP and elsewhere, they conclude that "we should generally enable actors to figure out through experimentation, feedback, and imitation, what choices best suit their interests . . . [while] policymakers should turn their attention to specific choice environments that remain resistant to the spontaneous learning of individuals despite their own ongoing interactions" (2021). To guide the latter process, they offer three principles they hope will keep such efforts within the bounds of liberal constitutional democracy. Those principles are *subsidiarity*, *proportionality*, and *scientific basis*.

We broadly agree with these three principles. But note that all three principles are fundamentally constraining rather than enabling: they place ostensible limits on government intervention. If C&T wish to make an enabling case for intervention, it must lie in the negative space of these principles: *if these three principles hold, then intervention is acceptable.* And that is where we must depart from them. While these principles are necessary, we do not think they are sufficient unless they are interpreted in very strict fashion.

By *subsidiarity*, C&T mean "that the governance solution to social problems should be established and implemented at the lowest feasible scale complementary to learning and voluntary imitation by other associations and jurisdictions" (2021). As C&T admit, this principle would usually push decisionmaking to the individual or small-group level. So what kind of government intervention do they believe clears this hurdle? They start with providing information, as governments can supply information that individuals may lack. We concede this point in the book, noting that the value of information was never doubt, even in the neoclassical perspective (EP, 417). Where we get off the bus, so to speak, is when informational "nudges" are intended as manipulative rather than informative. This happens when interventions are designed, not merely to supplement the individual's stock of knowledge, but to trigger a bias in

service of a specific outcome. And, we regret to say, this is where C&T go next, as they indicate support for nudges "to manipulate the choice environment to influence momentary decision-making such as to smoke or to buy cigarettes in cases where actors might not be mindful of this information" (2021). So we're no longer talking about whether the individual has information, but whether they are sufficiently *mindful* of the information they have. The best-known tool of this sort is a graphic image designed to make harms of an activity like smoking more "salient" to the consumer. How salient is salient enough? How do we know when the information has achieved optimal salience? In practice, the answer seems to be when the consumer has changed their behavior as the planner sees fit.

C&T further suggest that efforts like these are "not manipulative if they were introduced as part of a democratic discussion where smokers themselves were included and generally approved of it" (2021). We deny that an intrusive or coercive policy ceases to be so simply because a majority has approved of it. Non-consenting citizens are still subject to the policy. Moreover, C&T's position seems openly at odds with their own subsidiarity principle. It is entirely possible, for instance, for a democratic process to implement policies far above "the lowest feasible scale" for learning and voluntary imitation. C&T emphasize that the involvement and approval of smokers (or other targeted groups) is key to their argument – but smokers are not monolithic, and it is possible for some smokers to coerce other smokers. We worry that, even within a single discussion, subsidiarity has already proved far too permissive.

By *proportionality*, C&T mean "interventions should not exceed what is necessary to achieve a publicly approved goal" (2021). We have no argument against this principle *per se*, but its application depends crucially on the goal in question. As we observe in EP, behavioral paternalism in practice often devolves into targeting outcomes – more saving, less obesity, etc. – rather than the alleged goal of advancing people's subjective preferences. Simplistic goals that ignore subjective values may nevertheless be "publicly approved." To their credit, C&T add that interventions "should generally not be used just to improve social welfare from the policymaker's perspective, but primarily to make it easier for individuals to discover and pursue their own understanding of their welfare" (2021). This sounds good, but we are unsure which policies they think qualify. Do any of the specific policies mentioned actually do this? Do graphic images aid in self-discovery and self-understanding?

By *scientific basis*, C&T mean that proposed interventions should be supported by evidence of both actual harm and efficacy (2021). We agree, of course,[4] but with the caveat that "harm" and "efficacy" are both normatively loaded. Is the undeniable evidence that smoking and obesity harm people's

[4] And yet the scientific evidence for many behavioral policies is insufficient (EP, Chapter 6).

health enough to justify intervention? We say no, because true net harm is subjective; it depends on how the individual weighs the health harms against non-health benefits. Likewise, is a policy to be deemed effective because it successfully reduces those health harms? Again, no, because that could mean being effective at forcing people *not* to pursue their own subjective welfare.

To reiterate, we have no argument against subsidiarity, proportionality, and scientific basis. We simply do not think they are enough. In the final chapter of EP, we lay out a longer list of key distinctions to keep paternalism in check, including some mentioned earlier – self- versus other-imposed, invited versus uninvited, informative versus manipulative – as well as competitive versus monopolistic, voluntary versus coercive, and private versus public (EP, 414–420). These are the principles we would like to see guiding policy.

Hands, like C&T, largely accepts our conceptual approach. He concludes that governments should "quit trying to do paternalist policy that equates paternalism with successful utility maximization and instead have the government spend its resources on . . . *preventing harm to others*" (2021), which mostly means "the traditional subjects for governmental action: positive and negative externalities and public goods" (2021). We mostly agree. We probably differ on which other-regarding policies are justified, but that would be a much larger discussion.

Perhaps because Hands is eager to champion governments addressing "social problems," he seems motivated to downplay arguments in EP that rely on "more generic criticisms of almost any type of governmental policy," such as "the rational ignorance of voters, government failure, the biases of governmental policymakers, the influence of special interests, [and] slippery slope arguments about the growth of government" (2019). No doubt, many of our arguments in EP are not unique to paternalistic policy. But we endeavor to show how these broader arguments apply specifically, and often more crucially, to paternalism. For instance, it's true that knowledge problems afflict all manner of government policies, but paternalism opens up whole new realms of knowledge problem – such as trying to discern "true" preferences and "true" extents of bias hidden within people's minds. It's true that special interests have a baneful influence on all policymaking, but paternalistic policies seem especially susceptible to the "Baptist and Bootlegger" variety of rent-seeking (indeed, its name derives from the paternalist context of liquor regulation). Even as "new" paternalism endeavors to rely on behavioral science for support, politically it tends to become allied with the same moralistic impulses that led to the "old" paternalism. While slippery slopes occur in many areas of policy, paternalism is rife with characteristics – particularly gradients – that encourage such slopes. And so on.

In short, if Hands reads our book as making a generic case against all intervention, paternalistic or otherwise, that was not our intention. As we say in the book, some but not all of our arguments apply to non-paternalistic

interventions. That includes efforts to use behavioral "nudges" for other-regarding purposes.

Hands makes one additional, and somewhat guarded, point. Although he wants government to focus on social problems rather than paternalism, he also says that if government attempts the latter, it "should be directed towards making people healthier, live longer lives, and so forth," rather than "trying to increase individual preference satisfaction" (2021). Although we appreciate the honesty and straightforwardness of that approach, the problem is that such approaches tend to be heavy-handed and to disregard individual preference entirely. Despite our many disagreements with the behavioral paternalists, we believe they at least get the target right: if we intervene at all, it should be to make people better off *by their own lights*, not someone else's. Softer tools are generally a better way to way to do that. To be clear, we don't favor those tools, either. But if the alternative is harder and more intrusive policies to encourage "improvements in human physiology" (2021), we'll take the soft ones.

Hargreaves Heap (hereafter H-H) has a concern similar to that of Hands, as his reply article is heavily focused on arguments about fiscal externalities and other non-paternalist issues. He quotes extensively from a rather small section of the book – basically an aside – where we briefly venture outside the book's primary topic. We only included that section because "we have encountered the fiscal externalities argument often when discussing behavioral paternalism," so we thought it appropriate to include a "sketch [of] a response here" (EP, 430). We are happy to concede that our arguments in that short discussion are not dispositive, and moreover, that genuine externalities can justify government intervention in some cases.

However, we also believe H-H has misconstrued the structure of the book's broader argument. The misconstrual comes out when he says, "Politicians do not materially care whether an intervention is paternalistic, so R&W's demonstration that nudging is paternalistic passes most by" (2021), and again when he says, "For most politicians, therefore, an argument that 'nudges' are paternalistic misses the point. Politicians do not embrace 'nudging' because they think it is not paternalistic" (2021). In other words, H-H seems to think our argumentative strategy was as follows: (1) prove that nudges are paternalistic, and (2) expect that politicians will therefore stop doing nudges.

This was not our strategy at all. First, no part of our argument hinged on convincing anyone that *all* nudges are paternalistic – which we do not believe, as some nudges are intended to serve the interests of others.[5] Our focus was on paternalistic interventions, including paternalistic nudges and paternalist shoves, but excluding non-paternalistic nudges. Second, we did not expect that politicians would abandon policies simply because they are paternalistic.

[5]The paradigmatic example is instituting an "opt-out" system for organ donor status. This intervention is aimed at increasing organ donation, which benefits the recipients, not the donors.

We do not simply assume that paternalism is bad and everyone knows it. Our purpose in the book was to explain *why* paternalistic policies are unjustified, even when based on behavioral research.

Of course, politicians may simply disregard our arguments. This seems to be H-H's primary criticism: that our arguments, correct or not, won't make a difference to politicians who just want to get things done. They may refuse to listen, or shrug, or do "some eye-rolling directed at the ivory tower" (2021). We agree this is possible. Politicians lack strong incentives to act on the basis of principled arguments – a point we make explicitly. Frankly, we'd be surprised if any real politicians even read our book. Our main audience was our academic colleagues, policy analysts, and sophisticated members of the public, who we hope will exercise an indirect influence on politicians, regulators, and judges.

In any case, H-H offers an alternative strategy that he believes has a better chance of influencing politicians: "In short, politicians need to be persuaded to think differently about their objectives: the business of politics ought to be the choice of rules and not the achievement of specific outcomes for particular individuals" (2021). While we appreciate the Hayekian thrust of this position, we fail to see how H-H's rules-only principle is any more likely to persuade politicians than our anti-paternalism principle. If anything, he is offering an even more abstract and academic standard. If H-H is correct that politicians don't care about principles, only about accomplishing things for themselves and their constituents, then why should H-H's principle restrain them any more than ours?

In addition, it is unclear what the rules-only principle actually rules out. All rules work through their impact on specific individuals. Politicians are very talented at crafting policies phrased as general rules that nevertheless target the specific people they hope to benefit (e.g., "all farms in mid-Atlantic states engaged in the production of tobacco with an acreage of at least . . . "). For a rules-only principle to have any bite, the necessary level of abstractness of rules would need to be defined (see Whitman, 2009; Rizzo, 2021).

H-H believes his rules-only principle would rule out nudges. We are not so sure. H-H's claim is that rules must affect the constraints that all individuals face, whereas nudges are ostensibly designed *not* to change constraints, only the behaviors that some people select within them, and thus do not qualify as rules (2021). We would point out that every state-imposed nudge we know of affects the constraints of at least some individuals. For instance, a mandatory opt-out rule for savings constrains the behavior of all employers in how they handle their savings plans. Many other paternalistic interventions, such as sin taxes and cooling-off periods, directly alter constraints on people's decisions and would therefore pass muster as rules; we're unsure whether H-H would classify these as "nudges."

Elsewhere, H-H says that his rules-only principle rules out nudges because nudges are allegedly designed to affect the behavior of only some individuals,

specifically the irrational ones (2021), rather than all people. But don't all rules work this way? The rule against murder is only intended to deter would-be murderers; most people's choices are unaffected. Any given rule is likely to be "binding" only for some classes of people.

We agree with H-H's concern about the increasing "personalization" of politics, wherein policies are designed to help (or harm) relatively specific groups of individuals rather than focusing on the public as a whole (2021). However, that concern is largely orthogonal to our concerns in the book. Many paternalistic policies are championed as a means to help large swaths of the public, under the claim that we are all "Homer Simpson" to some degree and thus all need an assist. We believe those policies are mistaken, too, and even a well-defined rules-only principle would do nothing to stop them. We therefore offer our anti-paternalistic principle to accompany other worthwhile principles of governance – while recognizing that politicians may disregard them all.

Conclusions

Science advances through criticism, or as Karl Popper says, through errors and error correction. We hope we have not made many errors. Nevertheless, we are indebted to both the supportive and critical comments because they have helped sharpen our arguments, and in so doing have advanced an important discussion.

References

Cowen, N. and A. Trantidis. 2021. "Soft, interventionism: A Hayekian alternative to llibertarian paternalism". *Review of Behavioral Economics*. 8(3–4): 341–360.

Coyne, J. 2009. *Why Evolution Is True*. New York: Penguin Group.

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

Grüne-Yanoff, T. 2021. "Boosts: A remedy for Rizzo and Whitman's Panglossian fatalism". *Review of Behavioral Economics*. 8(3–4): 285–303.

Hands, D. W. 2021. "Libertarian paternalism: Making rational fools". *Review of Behavioral Economics*. 8(3–4): 305–326.

Hargreaves Heap, S. P. 2021. "The 'problem' is different and so is the 'solution'". *Review of Behavioral Economics*. 8(3–4): 327–340.

Hayek, F. A. 1967. "The theory of complex phenomena". *Studies in Philosophy, Politics, and Economics*: 22–42.

Hayek, F. A. 1973. *Law, Legislation and Liberty, Vol. 1: Rules and Order*. Chicago: University of Chicago Press.

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

Koppl, R. 2021. "Against Expertism". *Review of Behavioral Economics*. 8(3–4): 361–377.

Matson, E. W. and M. Dold. 2021. "The behavioral welfare economist in society: Considerations from David Hume". *Review of Behavioral Economics*. 8(3–4): 239–258.

Peart, S. J. 2021. "On making and remaking ourselves and others: Mill to Jevons and beyond on rationality, learning, and paternalism". *Review of Behavioral Economics*. 8(3–4): 221–237.

Rajagopalan, S. 2021. "Inclusive rationality: Struggle and aspiration". *Review of Behavioral Economics*. 8(3-4): 259–283.

Rizzo, M. J. 2019. "Inconsistency of not pathological". *Mind & Society*. 18(1): 77–85.

Rizzo, M. J. 2021. "Abstract rules for complex systems". *European Journal of Law and Economics*: First View, 1–19. URL: https://rdcu.be/cy6Fb.

Rizzo, M. J. and G. Whitman. 2019. *Escaping Paternalism: Rationality, Behavioral Economics, and Public Policy*. Cambridge Studies in Economics, Choice, and Society. Cambridge, United Kingdom; New York, NY: Cambridge University Press.

Whitman, D. G. 2009. "The rules of abstraction". *Review of Austrian Economics*. 22: 21–41.

Whitman, G. 2006. "Against the new paternalism: Internalities and the economics of self-control". *Policy Analysis*. 563: 1–16.

Whitman, G. 2021. "Austrian behavioral economics". *Journal of Institutional Economics*: First View, 1–18.