Irrationality and Immorality:

Exploring the Ethical Dimensions of Behavioral Public Policy

Alejandro Hortal

University of North Carolina Greensboro

Wake Forest University

Abstract

This paper critically explores the ethical dimensions of Behavioral Public Policy (BPP), a domain grounded in the understanding that human rationality is bounded and that this limitation often leads to behaviors deemed irrational. By applying the behavioral lens, which posits that people operate under bounded rationality, BPP aims to craft interventions that safeguard individuals against their biases. However, this approach raises significant ethical concerns, both in the scientific underpinnings of BPP and its application through policy interventions. Accordingly, this paper examines two distinct ethical dimensions of BPP, both as a scientific discipline and through its intervention methodologies. The analysis of the first dimension argues that, viewing bounded rationality as only instrumental in decision-making processes, risks oversimplifying ethical complexities. Such simplification may ignore the moral and value-based rationalities underlying decisions, potentially misattributing instances of immorality and akrasia (the failure to act according to one's judgment) to mere deficiencies in rational thinking. Secondly, the paper examines the impact of BPP on moral behavior and character development, addressing ethical concerns like the evidentiary basis of BPP research, the bounded rationality of choice architects themselves, and the morality of behavior changes induced by policy. Overall, this article provides a much-needed examination of the moral considerations associated with behavioral strategies in public policy and its epistemological foundations, contributing to a more comprehensive understanding of their ethical implications

Keywords: Behavioral Public Policy, Rationality, Bounded Rationality, Ethics, Morality

Introducción

Standard public policy and economics before the 1950s assumed humans were completely rational, consistent, egotistic, and utility maximizers. Daniel Kahneman explains that the belief that agents behave rationally (in this sense) is a foundational concept across many theories in the

social sciences (D Kahneman, 1994, p. 18). Those models failed to provide a robust explanation of people's decision-making. Herbert Simon, criticizing this idealistic approach, proposed a more empirical view of human rationality, bounded by its cognitive capacity, memory, reaction time, and the complexity of the environment. Simon believed that people do not optimize; they rather *satisfice* (1956). Following Simon's line of research (with essential differences that this paper will not examine), Kahneman himself together with Amos Tversky showed that our rationality is systematically and predictably limited (1974). When deciding, they argued, we tend to rely on heuristic processes which make us prone to biases. These biases, our limited cognitive capacities, and the complexity of the environment in which we make decisions increase the difficulty of behaving according to what we think is right and acting in a way that would improve our well-being. This separation from the normative rational standards can be frequently observed in people's behavior: even in cases where they have the knowledge, desire, and the means to do it, people often fail to eat healthily, save money, meet deadlines, or exercise.

Kahneman by himself (2003), or in collaboration with Tversky, Richard Thaler (Daniel Kahneman et al., 1986), or others, continued to explore the intersection between economics and psychology, highlighting how cognitive limitations and systematic biases affect economic decisions and personal well-being. Thaler, along with Cass Sunstein, following this line of research, applied insights from the field of behavioral economics to law (Jolls et al., 2000), and public policy issues (R. H. Thaler & Sunstein, 2003), and in 2008 they published the book *Nudge*. Although not the first one to do it, Kahneman had already wondered about a paternalistic approach in public policy since the bounded aspect of our rationality may impose deficiencies in the way we decide and how we think about our future, arguing that the government may have better and more objective knowledge about our well-being:

> The substantive question on which we focus here is whether choices maximize the (expected) utility of their consequences, as these consequences will actually be experienced. Accurate prediction of future tastes and accurate evaluation of past experiences emerge as critical elements of an individual's ability to maximize the experienced quality of his outcomes. Demonstrated deficiencies in the ability to predict future experiences and to learn from the past emerge as new challenges to the assumption of rationality. More provocatively, the observed deficiencies suggest the outline of a case in favor of some paternalistic interventions, when it is plausible that the state knows more about an individual's future tastes than the individual knows presently. The basis of these developments is an analysis of the concept of utility, which is introduced in the next section. How much do people know about their future tastes? Is it likely that an objective

> observer (or a government) could make more accurate predictions than individuals would make on their own behalf? (D Kahneman, 1994, p. 20)

Kahneman appears to question whether this superior (or more objective) knowledge may justify paternalistic interventions, given that it was deemed appropriate for Ulysses in his encounter with the sirens. Following this line of thought but with a libertarian twist, Thaler and Sunstein, assuming the bounded rationality of individuals, propose a new model for public policy-making based on a libertarian paternalistic perspective to decision architecture in their book *Nudge* (2009). Nudges are alterations in the choice environment in which we make decisions (a cafeteria, a webpage, a way of forming sentences and choosing certain words, the placement of a recycling bin in a room, etc.) that, considering the bounded reality of our rationality and decision-making capacities, seek to change the way we behave without affecting our freedom of choice and respecting our will. Under this approach, the decision architect will arrange the environment in such a way as to make certain options salient (or sometimes making them the default) to increase people's well-being, according to what the consensus of society (or the government, if we think in Kahneman's terms) defines as well-being (Hortal & Segoviano Contreras, 2023): better health, more savings, better education, environmental care, longer life expectancy, etc.

Improving well-being through nudges involves subtly guiding people towards better choices without coercing them, respecting their autonomy and capacity for self-determination. This approach is seen as a gentle and respectful way to influence behavior, contrasting with more invasive or paternalistic methods. The concept has gained traction not only in economics and public policy but also in health care, education, environmental policy, and personal finance, demonstrating its versatility and the wide-ranging potential for a positive impact on society. Although nudges are the most renowned and increasingly popular form of intervention, BPP encompasses a diverse array of tools inspired by behavioral sciences. This includes budges, boosts, shoves, and sludge audits, to name a few. What unifies these tools is their foundation in behavioral science research, which focuses on designing interventions that take into account people's bounded rationality and various biases.

Recent literature has extensively examined cognitive biases (Blumenthal-Barby & Krieger, 2015; Das & Teng, 1999; Haselton et al., 2015; Hilbert, 2012); however, with a few exceptions, moral biases have primarily been characterized as distortions of our moral behavior driven by self-interest (Croson & Konow, 2009). Nonetheless, there are instances where our

morality may be constrained by our rationality beyond self-interest. These are two different but connected cases, and reducing akrasia (the failure to act according to one's own principles) and immorality to irrationality (or bounded rationality) and vice versa can erroneously affect our understanding of people's rationality in all its approaches (cognitive, moral, instrumental, social, etc.). Although moral intuitions and decisions can be influenced by cognitive biases to the same extent as our economic or financial choices (Caviola et al., 2014), it is important for researchers not to equate morality solely with rationality. Our moral judgments are sometimes prone to biases, leading to actions that may not always align with our moral principles. This inconsistency can stem from complexities in decision-making, limitations in memory, and our inherently restricted and biased cognitive capabilities. However, this does not justify the conclusion that a deficiency in morality is necessarily due to a lack of rationality. This paper argues that while choice architecture, including nudges, can play a crucial role in addressing moral challenges, it cautions against reducing immorality and akrasia to simple outcomes of cognitive biases or irrational behavior due to our limited rationality.

Inevitably, any effort to influence human behavior warrants deep epistemological and ethical scrutiny, due to the potential risk of manipulating individual autonomy or causing undesirable effects. Such ethical analysis must also focus on the relevance of implementing these strategies within governmental entities. If, as postulated, these interventions have the capacity to enrich the lives of individuals (according to their own assessments of improvement) without compromising their freedom, then their more frequent use could be justified in order to promote a more robust well-being.

Although some of these interventions have proven successful, many of the studies backing them lack solid data to support their implementation or are based, at times, on trials that are not robust enough to extrapolate to a larger number of people or to carry them out in different contexts, times, or places. Sometimes, when the data yield favorable results, they are not adequately compared with alternative interventions that could be more effective. The epistemological problems in BPP are varied and complex, and the external validity of randomized controlled trials (RCTs) raises doubts about generalizing the results to real-world contexts. For example, a study on this topic (DellaVigna & Linos, 2020) claims that while the average impact of a nudge in academic articles is 8.7%, in institutions dedicated to BPP (also called Nudge Units), the impact is only 1.4%. Challenges in measuring and operationalizing behavioral changes can lead to biases and misinterpretations. The replicability crisis in

psychology and behavioral sciences has called into question the reliability of foundational studies. The complexity of human behavior, along with selection biases and the limitations of quantitative methods, makes it difficult to isolate the impact of interventions. Long-term effects and often underestimated unintended consequences are also concerning. Moreover, political and ideological influences may bias the choice and implementation of interventions, while interdisciplinary integration presents additional methodological challenges. These issues underscore the need for a careful and ethical approach in formulating and implementing policies based on behavioral sciences.

When influencing human behavior, ethical considerations for such interventions become necessary, as they can always endanger individuals' autonomy. Additionally, further ethical questions arise when behavioral interventions successfully modify people's conduct. It's worth questioning whether it's feasible to use such interventions for moral improvement or for strengthening virtues and individual character. This approach would not necessarily be a type of moral enhancement (Douglas, 2011; Shook, 2012), but a type of extended morality[1] (extended to the tools, phones, apps, choice environments we use). In the case of a nudge-type intervention, for example, inducing a change in a person's behavior, there's doubt as to whether such behavior can be considered morally superior (or inferior), despite having been stimulated unconsciously or impulsively, following the 'system-1' model proposed by Kahneman (2011). We should also not overllok an important issue that we must address: the theoretical framework of BPPs based on models of bounded rationality might excessively reduce moral richness to simple aspects of rationality, minimizing the importance of ethics inherent in the human decision-making process. That is, there's a risk of falling into a sort of Socratic moral intellectualism by thinking that people are not bad (or vicious, or selfish, or rude, for example), but are subject to biases that make them irrational. For Socrates, evil is caused by ignorance, perhaps some BPP approaches are tempted to think that immorality is mere irrationality, reducing the moral richness into mere aspects of rationality, thereby minimizing the significance of ethics inherent in the human decision-making process.

[1] Extended morality is a novel concept Alejandro Hortal is developing. Based on the notion of 'extended cognition', his thesis suggests that moral processes can extend beyond our brains to include external devices and environments that assist thought. Our ethical reasoning and moral behaviors can, therefore, be influenced, shaped, or even partially constituted by the technologies and tools we interact with. Instead of directly enhancing human capacities, this approach focuses on designing and utilizing external aids (e.g., smartphones, apps, AI systems) in ways that promote ethical thinking and moral actions.

The essay will, therefore, explore the ethical aspects of BPPs from various angles: assessing their effectiveness through existing evidence, examining how they impact individual autonomy, considering whether they can foster or weaken our morality, and questioning if their theoretical framework reduces morality to mere bounded rationality. The initial section explores the possibility of oversimplification of moral principles to mere rational calculations. This analysis seeks to unpack the nuances and limitations of equating morality strictly with rationality. Subsequently, the second section broadens the scope to critically assess the multifaceted implications of BPP interventions. It aims to examine the ethical tangles arising from insufficient empirical support, the questions of autonomy these interventions might challenge, their transparency, and the ramifications on decision-making processes due to the inherent limitations in the choice architect's rationality. The fundamental goal of the text will be to provide the reader with an introduction to some of the philosophical issues that emerge from the development of BPP. This approach aims to foster a deeper understanding of the intricate balance between the potential benefits of applying behavioral sciences to public policy and the essential need to navigate the ethical challenges they present.

The Irrationality and Immorality dimension: Socrates, moral intellectualism, and BPP

Rationality: a concept and an idea

The examination of rationality and its diverse implications across various disciplines offers a profound understanding of human behavior and morality. Spanish philosopher Gustavo Bueno's distinction between concepts and ideas (Bueno, 1993) can be useful to provide a framework for understanding various forms of rationality—economic, psychological, or moral—each concept associated with specific fields of study. While concepts are used within the different scientific fields, ideas are generally used beyond them. Accordingly, the concept of rationality in economics traditionally has dealt with efficient resource allocation to maximize needs and desires within the economic field. Rationality in the field of psychology generally involved individual thought processes and decision-making within that science, focusing on how information is processed and decisions are made (here we would have to distinguish between the different types of psychology -social, behavioral, etc.). Moral rationality concerns the ethical principles and judgments regarding right and wrong within the scope of ethics, which steers human actions toward what is deemed morally acceptable. Ethics, a subfield of philosophy, delves into the study of morality. These concepts span across diverse disciplines, much like how other concepts find relevance in various fields of science. For instance, 'freedom' pertains to Political Science when

discussing societal liberties, to Psychology in the context of free will, and even to Physics when referring to the movement freedom of electrons.[2] Ideas, according to this approach, transcend specific, finite, and sectorial fields, acting as a second-order construct that reflects upon and coordinates relationships between more defined 'concepts.' Ideas are not merely abstract or divine inspirations but are rooted in the connections established among concepts across different fields of knowledge. This conception positions ideas as reflective, integrative entities that draw from and go beyond the precise, delimited nature of concepts within technological, scientific, or even mythological contexts. Essentially, ideas in this framework serve as the bridges that link various domains of understanding, offering a platform for the synthesis and analysis of knowledge across disciplines. They embody a deeper level of cognitive engagement with the world, allowing for the examination of structures, relationships, and principles that transcend the immediate specificity of concepts. Ideas are shared in society, but they are also the field of Philosophy as a discipline.

Consequently, the idea of rationality and irrationality transcend the specific fields (Psychology, Economics, etc), mirroring a broader, reflective understanding of how these concepts interact with each other and with additional concepts on a second-order level. The idea of rationality, therefore, is not confined to a single field but overflows the boundaries of individual disciplines, offering a more integrative and reflective view on how rationality is understood and applied across various human situations. This interesting approach highlights the complexity of rationality as both a dissectible phenomenon into specific components (concepts used in different fields) and a broader, unifying dimension echoing the human capacity to reflect on and coordinate these diverse aspects of our understanding and action in the world. Rationality, therefore, can be seen as a concept (with different perspectives and ways to study it, as it is part of the different sciences or disciplines) or as an idea that transcends those disciplines.

The distinction between concepts and ideas elucidates the multifaceted challenges encountered across various disciplines in grappling with the concept of rationality. It simultaneously offers insights into the pitfalls of reductionism when applying rationality within discrete fields or as a broader idea. The idea of 'rationality' as articulated by ethicists diverges fundamentally from its conceptual interpretations in psychology or economics. Although these

[2] Electrons have three degrees of freedom in classical electrodynamics, six degrees of freedom in approximate models, and four degrees of freedom in quantum field theory

disciplines may reference the same underlying notion of rationality, the application and understanding of the concept are distinct, reproducing the variability observed with other concepts. For instance, the concept of 'love' manifests uniquely across disciplines such as Chemistry, Sociology, Biology, Theology, and Art. In each field, 'love' is studied, explored, and used contextualized with the conceptual elements that are related to that specific field in a manner that prevents its reduction to other domains. Specifically, 'love,' as a concept examined within Theology, cannot be solely interpreted through a psychological lens, without having to denied the existence of interpretations that traverse both fields without diminishing one in favor of the other. This approach highlights the importance of acknowledging the distinct contexts and applications of concepts like 'rationality' and 'love' across disciplines, thereby enriching our understanding of these ideas without succumbing to the limitations of reductionist thinking. Similarly, 'entropy', for instance, serves as an example of how a single idea can manifest distinctly across different scientific disciplines, reflecting the diverse contexts in which it is applied. Originally rooted in thermodynamics, entropy represents a measure of disorder or randomness within a system. In this classical physical context, it quantifies the amount of energy in a system that is not available for doing work, serving as a fundamental principle in understanding the direction of spontaneous processes. However, in information theory, entropy measures the unpredictability or the amount of information content in a message, essentially quantifying uncertainty or the degree of surprise associated with a particular set of outcomes. In ecology, entropy is employed to describe the diversity and stability of ecosystems, assesing the randomness in the distribution of species within an ecosystem, providing insights into its resilience and health. The idea of entropy would transcend those fields, although taking all of them into consideration, assuming that one approach to the concept cannot be reduced to the other one.

Interconnection, not reductionism

Some authors have explored the interconnection of moral and economic rationality from different angles (Baier, 1977; Nelson, 1988; Sahlin & Brännmark, 2013). The study of rational behavior in BPP often highlights concepts like bounded rationality, biases, and heuristics to explain our decisions, sometimes categorizing them as irrational. While BPP applies these concepts to explain the psychological and economic dimensions of rationality, its approach to ethics could oversimplify complex moral phenomena. This simplification risks portraying moral and immoral actions merely as issues of rationality, potentially minimizing the importance of individual moral accountability. By focusing solely on how biases affect decision-making, this

viewpoint might neglect the significance of moral values, our ability to act in alignment with these values, and the inherent responsibility of our human condition. This form of reductionism could mistakenly equate human moral conduct solely with rationality, whether in decision-making, psychology, or economics, leading to the mistaken belief that moral shortcomings are merely instances of irrational behavior. Such an outlook threatens to undermine the essence of morality, particularly regarding virtues like temperance or justice, by implying that immoral actions or akrasia are not truly unethical but are simply outcomes of limited rationality or irrationality. This dilutes the concept of morality, reducing it to a function of rationality rather than a complex interplay of values, decisions, and responsibilities.

In the field of BPP, there's a subtle but significant shift where researchers often move from discussing rationality in ethical terms to framing it within a psychological context. This transition from ethical to psychological or economic rationality is more implied than directly stated, suggesting a nuanced redirection of the discourse. Sahlin and Brännmark (2013), for example, use a psychological approach to examine the concept of rationality in their analysis. They draw on empirical findings from psychology to challenge the classical philosophical notions of rationality that underpin many traditional ethical theories. Similarly, other research assume that many cases of immorality or akrasia are caused by biases. Although in some instances this could be the case, it would be a mistake to asume that people are just irrational and never immoral. For example, the study 'Increasing altruistic and cooperative behaviour with simple moral nudges,' conducted by Valerio Capraro et al. (2019), employs a psychological approach to rationality in the context of ethical decision-making and pro-social behavior. This research examines how simple moral reminders, or 'nudges' influence individuals to make more altruistic and cooperative choices in various settings, including economic games and charity donations. The authors, therefore, use a concept of rationality from the psychological field to explore its application in ethics, particularly in encouraging pro-social behaviors through moral nudges. While these bridges are epistemologicaly viable, researchers have to acknowledge the differences in disciplines, and recognize that although there are some interconnections, ultimately each field uses a different approach to the concept of rationality.

Socratic moral intellectualism: A parallelism?

Socrates, at a different level, since instead of concepts he was connecting ideas with low empirical foundation, developed a theory similar to the combination of elements we explored in the previous part. According to Socrates, virtue is synonymous with knowledge: if an individual truly understands what is right, they will invariably choose to do it, implying that moral virtue and knowledge are inseparable. This perspective suggests that immoral actions result from a lack of understanding about what is genuinely good or beneficial, rather than a deliberate choice to do evil. Applying Socrates' moral intellectualism to a scenario where a person acts wrongly, not because of ignorance but due to bounded rationality (or irrationality) and cognitive biases, offers a modern twist. Here, the individual may possess knowledge of what is right but fails to act accordingly because cognitive limitations and biases distort their decision-making process. For instance, even if someone understands that saving for retirement is wise (knowledge), cognitive biases like present bias (overvaluing immediate rewards over future ones) may lead them to spend imprudently. In this context, Socrates' model could be expanded to argue that ethical action not only requires knowledge but also the ability to apply that knowledge free from the distortive effects of cognitive biases. This suggests a nuanced view where the cultivation of wisdom includes developing strategies to mitigate the impact of bounded rationality on ethical behavior, implying that moral education should also address understanding and overcoming our cognitive limitations. Holding this perspective in all cases can be epistemologically harmful since we would be neglecting our moral rationality, sometimes reducing it to mere biases. Some of those would be removed from the category of ethics and inserted in the category of epistemology. According to this risky approach, arrogance could be reduced to confirmation bias while incontinence (that wonderful Aristotelian word that means lack of self-control or discipline) can be a mere manifestation of present bias. Xenophobia would just be a type of in-group bias, and selfishness would stop being a moral problem since it can be classified as just a variety of self-serving bias.

Both, Socrates' approach and the second twist, propose a reductionistic perspective of our morality where ethics becomes psychology (or behavioral economics), but they both do it in different moments in history. While socrates deals at a philosophical level with ideas, BPP and the different scientific categories it uses are developed enough to use their own epistemological concepts independently. Although there are cases in which we can reduce some moral wrongdoings to biases and the types of cases studied by behavioral economics and BPP, it is important to categorize them with rigor. For example, in a reductionist view, a judge who does

not issue fair sentences because she is hungry (Kerry et al., 2019) would not be seen as morally questionable under this perspective, but simply as irrational or biased, which would reduce our morality to rationality, reflecting a view similar (only similar) to Socrates' moral intellectualism.

Against reductionsim

In concluding, it is crucial to underscore the inherent limitations and risks associated with the reductionist tendencies within BPP, especially when it attempts to distill the concept of moral rationality to mere psychological phenomena. This approach, while offering valuable insights into human behavior, risks oversimplifying the complex landscape of ethical decision-making and moral responsibility. Max Webber (1978) claimed that rationality was a complex idea that cannot be reduced to an instrumental conception, defending that it manifests itself as expressive, social, cognitive, etc. Raymond Boudon too argued that to be rational, people must have reasons to act in a specific manner, separating himself from any reductionist perspective that would assume that rationality is just instrumental. In a similar approach, recently some authors (Echeverría & Álvarez, 2008; Hortal, 2020) have argued that rationality is not only instrumental, it is also bounded and axiological. Rationality, therefore, can be examined in different ways and different disciplines will do it from diverse irreductible perspectives using methodologies adscribed to their own field.

The reduction of moral rationality to psychological or economic rationality not only undermines the intricacy of ethical considerations but also potentially erodes the moral agency. By potentially attributing immoral or non-pro-social actions merely to cognitive biases or bounded rationality, such an approach risks absolving individuals of responsibility for their actions. It paints a picture of humans as mere products of their psychological limitations, rather than as agents capable of moral growth and ethical reflection. We could end up with a-moral agents. Furthermore, by focusing predominantly on modifying behavior through psychological insights, BPP inadvertently marginalizes the role of ethical education and moral reasoning. The development of virtue, understanding of ethical principles, and cultivation of moral wisdom are relegated to secondary importance behind the manipulation of decision environments. This stance, while not denying the utility of psychological insights in promoting certain behaviors, fails to appreciate the full breadth of what it means to be a moral agent.

In essence, while BPP offers valuable tools for influencing behavior and promoting societal well-being, its reductionist tendencies towards moral rationality pose significant ethical and philosophical challenges. A more holistic approach, acknowledging the integral role of moral reasoning, ethical deliberation, and the cultivation of virtues, is imperative. Only by embracing the complexity of human morality and the diverse influences on ethical decision-making can policies truly foster a society that values and upholds the moral dignity and autonomy of its members.

There are other issues besides those found at the border between ethics and epistemology that are also related to the theoretical framework of BPP. Ethical analysis can also explore whether the influence of nudges (or other behavioral interventions) on a person's decision-making, such as facilitating certain choices, prevents considering those final actions as morally good (as sometimes we do not see opposite cases as morally wrong). For instance, consider a scenario where an individual is motivated to reduce energy consumption at home, thus benefiting the environment and practicing moderation, due to a billing system that leverages their competitive nature by comparing their energy usage with that of their neighbors. Does the external motivation of competition diminish the moral value of their action? Similarly, in the case of organ donation, if the default option on registration forms is set to encourage donation, leading more people to become donors because opting out requires extra effort, does this external influence undermine the moral worth of their decision to donate? This discussion prompts us to question whether actions influenced by such policies should be viewed merely as the outcomes of those policies, rather than as expressions of individual moral virtue.

The ethics of BPP's interventions

Initial considerations

In the seminal article that originated the ethical analysis on nudges, written by Luc Bovens (2009), it is found that the use of nudges can be justified to correct 'ignorance', where there is a lack of specialized knowledge, such as in decisions about retirement plans or medical treatments. Also, according to Bovens, they can be justified to combat 'inertia', in situations where knowledge is present but procrastination prevents action. Nudges can be used to reduce 'akrasia' or weakness of will, as in the cases of excessive consumption of unhealthy foods or lack of savings, and can also intervene when there is 'repugnance' in situations of emotional cost, such

as in the decision to become an organ donor, and in cases where there is 'exception', that is, where some decisions may cause regret, though this does not apply to all subgroups (gender change, buying an expensive car, etc.). Lastly, nudges can also seek 'social benefits', balancing individual decisions that are not socially beneficial, as in the case of environmental issues. Although the use of nudges can be justified in these types of situations, there are other angles within their ethical analysis that we must consider.

An intriguing concept to explore is the potential of dynamic choice architecture—characterized by its adaptability through algorithms and AI to suit the evolving preferences and contexts of individuals—to serve as a form of extended morality. This notion parallels the theory of extended cognition, which posits that external devices (like smartphones, applications, smartwatches, and other digital tools) can be seen as integral extensions of our cognitive processes. In a similar vein, we could conceive a theory of extended morality, where these technological aids are not merely external artifacts but are woven into the fabric of our moral being. As these devices guide and influence our choices and actions, they could effectively become external embodiments of our moral compass, shaping our ethical decisions and behaviors in a continuously interconnected world. This raises profound questions about the extent to which technology can and should play a role in the moral dimensions of our lives, potentially redefining the boundaries of moral agency and responsibility. These tools, functioning as dynamic choice environments that adapt and update in response to varying contexts, have the potential to serve as external enhancers of our moral character. Unlike the concerns highlighted by Bovens regarding the necessity of such aids stemming from specific deficiencies or issues, the acceptance and utilization of these tools can be motivated by a simple recognition of the desire for additional support in navigating our moral landscape. This perspective emphasizes a proactive approach to moral enhancement, where individuals seek out these technologies not as a remedy for shortcomings but as a means to augment and refine their ethical decision-making processes. Through this lens, the use of such dynamic, algorithm-driven aids transcends the corrective framework and ventures into the realm of moral optimization, offering a contemporary avenue for individuals to expand and enrich their moral capabilities.

Lack of evidence as an ethical problem

The multivariability and complexity of the environment and human behavior can cause long-term effects and unintended consequences. Many studies on nudges and other behavioral

tools focus on immediate outcomes without considering the long-term effects. Sometimes, the sheer impossibility of waiting due to academic work demands that compel researchers to 'publish or perish' motivates them to release articles and research findings as soon as positive immediate results are obtained. This leads to biases in data selection and a neglect of measuring the long-term consequences of the policies under study, increasing the likelihood of unintended consequences that only become evident over time. Ethically, this is problematic.

The application of behavioral tools in public policy can be influenced by political or ideological biases, affecting which interventions are chosen and how they are implemented. In some cases, rather than using evidence-based policies, evidence is selected to support pre-existing political agendas. Additionally, an overreliance on quantitative methods like RCTs may overlook qualitative aspects of human behavior, such as motivations, beliefs, and cultural factors, which are crucial for fully understanding the impact of interventions or the reasons behind certain behaviors. Qualitative methods can elucidate the causal mechanisms of potential interventions and behavioral changes. However, even if a behavioral intervention is successful, it's important to consider how this change contributes to the ultimate goal sought, which may not rely solely on behavioral changes but also on systemic alterations. For instance, as previously mentioned, increasing the number of organ donors does not necessarily increase the number of organ transplants. Behavioral interventions that lead to a higher number of donors must ensure that this translates into what is fundamentally sought: more transplants. All these issues can also represent moral challenges.

The studies conducted on the ethics of BPP, in general, and nudges, in particular, have mainly focused on examining whether such interventions infringe on people's autonomy (Vugts et al., 2020), as they often target System-1 (unconscious, automatic). However, autonomy, whether as freedom of choice, agency, or self-constitution, is not the only issue. Any ethical analysis of BPP must be approached from multiple perspectives, given that they are composed of intricate elements with potential moral relevance in many aspects, from the legitimacy of their use to the possibility of reducing immoral behavior to mere irrationality of the subjects, potentially making morality disappear in many of our actions and behaviors within their theoretical framework. Such analysis, for ethical reasons, must also refer to the empirical bases that support BPP and whether they make them intrinsically more effective compared to other methods. These epistemic deficiencies can cause ethical problems of legitimizing the

interventions used. Likewise, in addition to comparative analysis, it is crucial to evaluate whether a specific intervention has been empirically validated for application in a particular population. The lack of evidence (whether comparative or not), the lack of robustness of the studies, or biases in research on a general topic or on a particular intervention, can generate criticisms of their effectiveness and applicability. Therefore, any measure seeking to change behaviors in society must be based on solid and demonstrable research.

One of the most comprehensive meta-analyses conducted recently (Mertens et al., 2022), after analyzing over 200 studies (with more than 2.1 million participants), found that these generally have a small positive effect on behavior change (Cohen's d = 0.43). This is comparable to more traditional intervention approaches like educational campaigns or financial incentives. The study also confirms that effectiveness varies according to the technique and domain, with interventions focused on decision structure being more effective than those centered on information or decision assistance. The study notes a moderate bias towards positive results in publication and discusses implications for theory and the formulation of behavior-based policies. Overall, the authors advocate for the effectiveness of choice architecture in changing behaviors across various contexts, populations, and places, highlighting its versatility and usefulness despite criticisms of its efficacy. They argue that its benefit is not limited to punctual interventions but also enhances hybrid policies, including economic incentives.

When addressing the ethical aspects of these interventions, it's recognized that while they promise positive outcomes, they are not a universal solution for behavioral change. The need to base any public policy on solid empirical research is crucial, especially to avoid adverse effects. However, the field is marked by variable quality research, with some studies offering statistically insignificant results or, in extreme cases, fraudulent practices, such as the accusations of data falsification in research conducted by Francesca Gino and Dan Ariely (Simonsohn et al., 2021).

In conclusion, interdisciplinary integration is a fundamental part of the field of BPP. This field, almost by definition, requires the integration of knowledge from psychology, economics, sociology, and other disciplines, and this diversity can lead to epistemological challenges, as different disciplines have different standards and methodologies for study. The issues we have detailed in this section highlight the complexity and challenges in the design, implementation,

and evaluation of behavioral interventions in public policy, suggesting a need for careful consideration, ethical reflection, and methodological rigor in the field.

The Bounded Rationality of the Choice Architect, Autonomy, and Transparency

The theoretical framework underlying all types of behavioral interventions is centered on the fundamental idea that our rationality is limited; some even claim that we are systematically and predictably irrational (Ariely, 2008). This means that the researcher, the public policy expert, or the decision architect is also a person with bounded rationality (or irrational), as are the editors and reviewers of the journals in which we publish our studies, or the people who participate in the studies or surveys we conduct. This leads us back to the previous ethical consideration: any intervention must be justified with robust studies. Nevertheless, to protect ourselves from the bounded rationality and biases of these agents organizing our lives, establishing a process of open and democratic review and debate (not limited only to parliaments) about the measures to be implemented is crucial. This ensures that they are transparent and clearly communicated, as this does not imply a reduction in effectiveness. For example, an interesting study (Bruns et al., 2018) explored the effectiveness and ethics of nudges, focusing on whether they can be transparent without losing their effect, showing that transparency, whether about the nudge's influence, its purpose, or both, did not decrease their efficacy. Moreover, psychological backlash was not found to affect the results. This finding is crucial for public policy, suggesting that it is possible to implement nudges in a transparent and effective manner, challenging the notion that they need to be subtle and covert to work. However, this approach poses a challenge for the responsible adoption of BPP based on decision architecture. Often, there may be a strong inclination to employ these strategies without solid evidence, without transparency, or without having had an appropriate social debate about their goals and the behavioral changes they seek, which can be problematic. Therefore, establishing empirical evidence, transparency, and social debate as prerequisites before implementing these policies is essential, connecting the demonstrated effectiveness of nudges with the need for an ethical and transparent framework in their application.

Altering Moral Behavior, values, and character

BPP might not only improve people's behavior but also their morality and virtue. Thus, in the context of behavioral economics and public policy, it's possible to influence the formation of

virtuous habits through nudge-type interventions. These interventions can be applied in various areas, such as public health, environmental concerns, and retirement savings, promoting the development of morally virtuous habits and encouraging deliberation. The implementation of what Hortal (2024) has denominated virtue nudges aims to align habits with ethical virtues to modify behavior. In his article, Hortal highlights the transformative potential of these subtle nudges in the formation of individual character through habit formation. Designed to catalyze positive behavioral change, nudges encourage the development of virtuous habits and the internalization of these tendencies by individuals, contributing to long-term character development. Thus, a new categorization within behavioral public policy emerges, known as virtue nudges. These nudges respect the individual's will and promote the formation and maintenance of habits associated with virtuous behaviors. Their goal is to help people cultivate or maintain virtuous behaviors, offering a new perspective to understand the ethical implications and social impact of choice architecture and BPP. Virtue nudges are designed to be effective and respect personal choice, with the aim of fostering virtuous lifestyles. It's important to emphasize the significance of integrating these nudges, along with other interventions, into public policies to promote and encourage virtuous behaviors in society.

Conclusión

In conclusion, the aim of this paper has been to bridge the gap between the scientific underpinnings of BPP and its ethical implications. This analysis reveals a complex interplay between human rationality, ethical decision-making, and the role of public policy in navigating these realms. By critically examining both the scientific discipline of BPP and its applied methodologies, the paper underscores the necessity of integrating ethical considerations into the fabric of policy design, implementation, and its epistemological foundations.

This work called attention to the risks of oversimplifying human behavior as merely a product of bounded rationality, thereby potentially overlooking the deeper moral and ethical dimensions that underpin decision-making processes. It argues for a more nuanced approach that recognizes the value of moral rationality, the development of virtues, and the importance of ethical education alongside behavioral interventions. The manuscript also examined the ethical dimension of the implementation of BPP interventions, from their unintended consequences, to their affect on autonomy and the possible negative effects of the choice architect’s bounded rationality.It also highlighted positive aspects of how to use BPP and its tools in the development

of morality in individuals, exploring how dynamic choice architecture in technology can be effectively used for moral enhancement, and how virtue nudges can be employed to develop moral character through habituation.

Looking ahead, future research should strive to further clarify the epistemological contours of BPP as a discipline, investigating not only its field of applicability and the concepts that belong to its scientific domains, but also how policies can be designed and applied in ways that respect and promote individual autonomy, moral agency, and societal well-being. This includes exploring the long-term impacts of behavioral interventions on character development, the evidentiary basis of BPP research, and the ethical considerations surrounding the choice architecture. Moreover, fostering transparency and engaging in democratic deliberation over the implementation of BPP interventions emerge as crucial steps toward ensuring that such policies align with societal values and ethical principles.

Ultimately, this analysis should serve as a foundational step toward reimagining the ethical landscape of behavioral public policy, highlighting the critical need to think about rationality from different angles, urging policymakers, researchers, and the broader public to consider not just the cognitive limitations of human behavior but also its irreductible moral dimensions. As we move forward, it is imperative that we embrace a view of rationality that goes beyond the one that emanates from behavioral economics, one that understand the similarities and, above all, the differences between moral, psychological, social, or economic rationality, to fostering a more just, virtuous, and rational society.

Bibliography

Ariely, D. (2008). *Predictably irrational*. Harper Collins.

Baier, K. (1977). Rationality and morality. *Erkenntnis*, *11*(1), 197–223. https://doi.org/10.1007/BF00169852

Blumenthal-Barby, J. S., & Krieger, H. (2015). Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. *Medical Decision Making*, *35*(4), 539–557. https://doi.org/10.1177/0272989X14547740

Bovens, L. (2009). The Ethics of Nudge. In T. Grüne-Yanoff & S. O. Hansson (Eds.), *Preference Change* (Vol. 42, pp. 207–219). Springer Netherlands.

Bruns, H., Kantorowicz-Reznichenko, E., Klement, K., Jonsson, M. L., & Rahali, B. (2018). Can nudges be transparent and yet effective? *Journal Of Economic Psychology*, *65*(33), 41–59.

https://doi.org/10.1016/j.joep.2018.02.002

Bueno, G. (1993). *Teoría del cierre categorial* (Vol. 1). Pentalfa.

Capraro, V., Jagfeld, G., Klein, R., Mul, M., & de Pol, I. van. (2019). Increasing altruistic and cooperative behaviour with simple moral nudges. *Scientific Reports*, *9*(1), 11880. https://doi.org/10.1038/s41598-019-48094-4

Caviola, L., Mannino, A., Savulescu, J., & Faulmüller, N. (2014). Cognitive biases can affect moral intuitions about cognitive enhancement. *Frontiers in Systems Neuroscience*, *8*, 195. https://doi.org/10.3389/fnsys.2014.00195

Croson, R., & Konow, J. (2009). Social preferences and moral biases. *Journal of Economic Behavior & Organization*, *69*(3), 201–212. https://doi.org/10.1016/j.jebo.2008.10.007

Das, T. K., & Teng, B.-S. (1999). Cognitive biases and strategic decision processes: an integrative perspective. *Journal of Management Studies*, *36*(6), 757–778. https://doi.org/10.1111/1467-6486.00157

DellaVigna, S., & Linos, E. (2020). RCTs to Scale: Comprehensive Evidence from Two Nudge Units. *RCTs to Scale: Comprehensive Evidence from Two Nudge Units*.

Douglas, T. (2011). Moral enhancement. *Enhancing Human Capacities*, 465–485.

Echeverría, J., & Álvarez, J. F. (2008). Bounded Rationality in Social Sciences. In E. Agazzi (Ed.), *Epistemology and the Social* (pp. 173–191). Rodopi.

Haselton, M. G., Nettle, D., & Andrews, P. W. (2015). The evolution of cognitive bias. *The Handbook of Evolutionary Psychology*, 724–746.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological Bulletin*, *138*(2), 211–237. https://doi.org/10.1037/a0025940

Hortal, A., & Segoviano Contreras, L. E. (2023). Behavioral Public Policy and Well-Being: Towards a Normative Demarcation of Nudges and Sludges . *Review of Behavioral Economics*.

Hortal, A. (2020). Nudging and Educating: Bounded Axiological Rationality in Behavioral Insights. *Behavioural Public Policy*, *4*(3), 292–315. https://doi.org/10.1017/bpp.2019.2

Hortal, A. (2024). Thriving by Design: Can Behavioral Economics and Public Policy Shape Virtuous Lives? *Behanomics*, *2*(1).

Jolls, C., Sunstein, C. R., & Thaler, R. H. (2000). A behavioral approach to law and economics. In C. R. Sunstein (Ed.), *Behavioral law and economics* (pp. 13–58). Cambridge University

Press. https://doi.org/10.1017/CBO9781139175197.002

Kahneman, Daniel, Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of Business (Chicago, Ill.)*, *59*(S4), S285. https://doi.org/10.1086/296367

Kahneman, Daniel. (2003). A psychological perspective on economics. *American Economic Review*, *93*(2), 162–168. https://doi.org/10.1257/000282803321946985

Kahneman, Daniel. (2011). *Thinking, Fast and Slow* (p. 511). Macmillan.

Kahneman, D. (1994). New Challenges to the Rationality Assumption. *Journal of Institutional and Theoretical Economics (JITE)*, *150*(1), 18–36.

Kerry, N., Loria, R. N., & Murray, D. R. (2019). Gluttons for punishment? experimentally induced hunger unexpectedly reduces harshness of suggested punishments. *Adaptive Human Behavior and Physiology*, *5*(4), 352–370. https://doi.org/10.1007/s40750-019-00121-4

Mertens, S., Herberz, M., Hahnel, U. J. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(1). https://doi.org/10.1073/pnas.2107346118

Nelson, A. (1988). Economic Rationality and Morality. *Philosophy & Public Affairs*, *17*(2), 149–166.

Sahlin, N. E., & Brännmark, J. (2013). How can we be moral when we are so irrational?. *Logique et Analyse*, 101–126.

Shook, J. R. (2012). Neuroethics and the possible types of moral enhancement. *AJOB Neuroscience*, *3*(4), 3–14. https://doi.org/10.1080/21507740.2012.712602

Simonsohn, U., Nelson, L., & Simmons, J. (2021). Evidence of fraud in an influential field experiment about dishonesty. *Data Colada*, *98*.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

Thaler, R., & Sunstein, C. (2009). *Nudge: Improving Decisions About Health, Wealth and Happiness*. Penguin.

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian Paternalism. *American Economic Review*, *93*(2), 175–179. https://doi.org/10.1257/000282803321947001

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vugts, A., Van Den Hoven, M., De Vet, E., & Verweij, M. (2020). How autonomy is understood in discussions on the ethics of nudging. *Behavioural Public Policy*, *4*(1), 108–123. https://doi.org/10.1017/bpp.2018.5

Weber, M., 1864-1920. (1978). *Economy and society: an outline of interpretative sociology* (p. 1470). University Of California Press.