Behavioural Public Policy (2023), 7: 2, 333–352
© The Author(s), 2020. Published by Cambridge University Press doi:10.1017/bpp.2020.55 First published online 30 October 2020

Evidence-based policies, nudge theory and Nancy Cartwright: a search for causal principles

ALEJANDRO HORTAL*
Department of Languages, Literatures, and Cultures, University of North Carolina Greensboro, NC, USA

Abstract: Nancy Cartwright argues that evidence-based policies should not only rely on randomized controlled trials (RCTs) to test their effectiveness – they should also use horizontal and vertical searches to find support factors and causal principles that help define how those policies work. This paper aims at analyzing Cartwright's epistemology regarding evidence-based policies and their use of RCTs while applying her findings to current research involving nudges as behavioral public policy interventions. Holding a narrowly instrumental view of rationality, nudge theory tends to neglect other expressive components. Policymakers, in their quest for causal principles, should consider the expressive rationality of individuals in their research. This inclusion would not only increase the effectiveness of nudges, but also address some ethical issues related to people's autonomy when targeted by these interventions.

Introduction

Since the publication of *Nudge* (Thaler & Sunstein, 2009), public policymakers have been using nudges as interventions based on behavioral insights to modify the behavior of citizens predictably by manipulating their choice environment. Governments and other institutions have created a multitude of nudge units that draw from this theory since its premise assumes the preservation of liberty (nudges do not coerce) while paternalistically it is capable of orienting behavior to the desired target (libertarian paternalism). According to Adam Oliver, nudges, as they focus upon internalities, are "the dominant framework in behavioral public policy to date" (2019, p. 147)

* Correspondence to: Department of Languages, Literatures, and Cultures, University of North Carolina Greensboro, Moore Humanities Research Building, Office 1119, Greensboro, NC. USA. Email: a_hortal@uncg.edu

Nancy Cartwright (2011b, 2012, 2013, 2018) claims that the methodology currently used in public policy to justify specific interventions in a particular place based on randomized controlled trials (RCTs) does not provide enough evidence to support the notion that the policy will work in a different setting. Using Cartwright's criticism as a philosophical frame, this paper aims at providing an epistemological analysis of current behavioral public policy interventions based on nudges (Thaler & Sunstein, 2009). These small changes in the choice architecture of a decision scenario promise a determined outcome that will result in an expected behavioral change of individuals without the need to impose bans or without the expenditure to educate people. According to their proponents, nudges tend to be effective because they are built on robust empirical experiments, most of them based on RCTs. This paper predominantly examines Cartwright's epistemology regarding evidence-based policies to shed some light on current advances in nudge theory. The core of the following pages will focus on her work on RCTs and her epistemological criticism, which provides a pertinent theoretical frame to philosophically understand nudges as libertarian paternalistic interventions. To do so, this paper starts by discussing Cartwright's philosophy regarding evidence-based policies and her theories on the epistemological status of RCTs. It will proceed by introducing the notion of nudges as the public policy representation of the psychological work carried by Amos Tversky and Daniel Kahneman and their view of human rationality as fundamentally and systematically flawed. Although Tversky and Kahneman cannot be classified as nudge theorists themselves, Thaler and Sunstein (2009) used their research as the theoretical foundation of the theory: their approach to heuristics and biases and the two-system model to understand cognition are broadly used by nudge theorists. The article will then consider how such interventions, while based on evidence, assumed a black box in the reasons for human behavior. Nudges mainly target the outcome of behavior – they are used instrumentally by public policymakers to alter people's conduct. To justify their deployment, they heavily rely on RCTs: "Nudging seems to be firmly positioned in evidence-based policy rhetoric, and encourages the use of Randomized Control Trials to determine the effectiveness of a policy" (Einfeld, 2019, p. 509). The UK's Behavioural Insights Team, the leading organization that uses behavioral insights in public policy (nudges), promotes the use of RCTs as the gold standard to evaluate nudges as interventions (Halpern, 2013).

Cartwright's epistemological theory will be used to analyze nudges and the reasons why they sometimes fail in their attempt to produce the desired behavioral change. Nudges are built on universal assumptions of human rationality, and they isolate human decision-making as their reference point. In order for policies to work in more places than that originally targeted by the RCT, as

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

Cartwright claims, we have to move beyond these types of trials and attempt to find causal principles and support factors. If policymakers expect nudges to work beyond their original target place and time, they must find those causal principles and support factors through what Cartwright called *vertical* and *horizontal* searches. When individuals behave, they do so to arrive at the desired end with their action, but they also act due to specific reasons: they have reasons based on preferences for doing what they are doing that go beyond the effect of the action itself. Understanding those reasons (our expressive and social rationality) should be a part of the research conducted by policymakers, since such reasons act as causal principles for our behavior and may affect the effectiveness of policies based on nudges.

Accordingly, our rationality is not only instrumental (goal-oriented), it is also social (intersubjective, based on social norms) and expressive:

> Expressive rationality seeks good reasons for individual action that refer to someone's enthusiasm, desires, feelings of sympathy and opposite emotions like fear, dislike or anger as motives … Decisions that require predominantly expressive rationality include decisions such as what career steps we want to make, how to decorate our house, to whom we want to get married, etc. Expressive rationality gives us orientation for such personal decisions, and also helps to explain to others why we made such decisions based on what was important to us. It relates to taste, well-being, emotional fulfillment and aesthetic considerations. (Bouwmeester, 2017, p. 45)

Tversky and Kahneman's research highlighted the irrationality of human behavior, since our heuristic cognitive processes often lead us to act against our rationality. Taking this into consideration, nudges are interventions that shape the choice environment to change people's behavior predictably, so they can get closer to the rationality they seek. By altering this choice environment, they make the rational choice the easiest choice, testing the efficiency and efficacy of policies using RCTs. Nudge theory, therefore, assumes an instrumental notion of rationality, not only because it conceives of the expressive components as either irrelevant or as a source for our irrational behavior, but above all because it presupposes the existence of a standard normative rationality that can be used as a rational criterion for optimization. Nudges dismiss, therefore, bounded rationality and bring back instrumental rationality, the one in charge of selecting the most efficient means to achieve our goals: "The focus of instrumental rationality is on how we can do things better, not on the goals people want to achieve. It excludes the question of why we think certain objectives or values are worth aiming at. Instrumental rationality thus ignores many good reasons to act on" (Bouwmeester, 2017, p. 5). Neglecting the expressive and social elements of our rationality may

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

cause public policymakers (e.g., nudge theorists) to consider specific actions as irrational when they are not selecting the most effective means for our goals. Nudge theorists could see these behaviors as incomprehensible and, therefore, a possible target of an intervention.

> [I]ndividuals derive "expressive utility," intrinsic and instrumental, from actions that, against the background of social norms, convey their defining group commitments … Identity-protective cognition is the style of reasoning for rationally engaging information that is relevant to identity-expressive beliefs, particularly when that information has no other real relevance to an individual's life. (Kahan, 2017, p. 28)

Although Cristina Bicchieri (2017; Bicchieri & Dimant, 2019) has been leading the effort of introducing social rationality in nudges to make the field more robust, I claim that nudge theory would benefit by considering a more comprehensive approach to rationality and behavior, one that includes our expressive rationality, among others. This paper, therefore, will also introduce the expressive components of rationality in nudge theory by conducting a critical approximation of the epistemological status of these interventions based on the work Nancy Cartwright has done on evidence-based policies and RCTs. Her claim to go beyond RCTs and find causal principles and support factors to ensure the effectiveness of policies is used to analyze nudge theory (since it heavily relies on RCTs), and it is complemented with a more comprehensive view of rationality. While Cartwright asserts that policies should find the reasons for behavioral changes if they want to be effective, this paper concludes that they must also understand the reasons for the initial behavior, and those reasons cannot just be reduced to biases: sometimes they are connected to the expressive or social components of our rationality.

Cartwright: what worked there may not work here

In 2012, Cartwright and Hardie published a philosophical study questioning how evidence-based policies rely on RCTs to export policies that have worked in a specific place and particular moment to another place in a different time. Cartwright explains that for evidence-based policies, RCTs are the gold standard for establishing what works (Cartwright, 2012, p. 298). The main issue, she claims, is that the most a good RCT can prove is that it worked somewhere, not that it will work in a different setting. Building upon that criticism, she provides a theory of evidence to evaluate the claims of RCTs regarding their possible applicability to other settings. She argues that evidence that the policy worked somewhere is only a good starting point. One also needs evidence that the policy will work in the target setting. The relevant question is whether the

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

policy will "make a positive difference in the desired outcome" (Cartwright & Hardie, 2012, p. 5).

RCTs are considered by the industry to be the most rigorous tool for predicting the *effectiveness* of a policy: "A well-designed randomized experiment makes it highly likely that the effect of the treatment be reflected in the data, but does not guarantee that this is going to be the case" (Guala, 2012, p. 615). Cartwright posits that, although policymakers would like to have evidence that those policies will work in other settings, RCTs cannot provide this – they can only inform on the *efficacy* of a treatment: that the policy worked somewhere (Cartwright & Hardie, 2012, p. 9); that is, "that it worked in the studied situation" (Cartwright, 2012, p. 299). To check for effectiveness, experts need to understand some facts about the causal role that the policy plays and what type of support factors it needs.

RCTs have a clear advantage: by separating control and treatment groups and creating a blind and random process of selection, experimenters can isolate the effects of specific policies on a group of individuals. But the process warrants some "metaphysical premises":

> RCT logic assumes a general metaphysical premise (premise 1) that probabilistic dependence calls for causal explanation. Experimental design acts to ensure premise 2: all features causally relevant to the outcome other than the treatment (and its downstream effects) are distributed identically between treatment and control groups. If the outcome is more likely in the treatment than the control group, which is premise 3, the only explanation possible is that the treatment caused the outcome in some members of that group. (Cartwright, 2011a, p. 1400)

RCTs, therefore, try to convert probability to causal explanation and to establish that if we observe a positive effect in the treatment group, it had to be caused by the policy. But, according to Cartwright, the outcomes of an RCT can only be applied in settings that share the same acting principles that produced the effect in the study situation (Cartwright, 2012, p. 313). Since causal principles are local, we should not expect that the policy will work somewhere else (Cartwright, 2012, p. 310). To have evidence that the policy will work somewhere else, we need to find causal principles climbing up the "ladder of abstraction" (Cartwright, 2012, p. 311), using more deliberative and thoughtful tools with the goal of finding a causal claim with the capacity to establish that the "treatment reliably promotes the outcome" (Cartwright, 2011a, p. 1401); that is, that the treatment is able to produce more cases of the outcome in a variety of circumstances.

Accordingly, Cartwright emphasizes the notion of *capacity*, considering it a "powerful tool," since that is what may show the effectiveness of the policy and,

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

consequently, that it will work somewhere else: “Effectiveness is what a cause does ‘in the field’” (Cartwright, 2009, p. 185). Policymakers, she states (Cartwright, 2009, p. 203), usually have good reasons to believe that their interventions will have an effect in part of the population before they test it with RCTs. They have theories, and although theories sometimes are uncertain, learning how to deal with those uncertainties should be a part of the task of finding the evidence that a policy will work somewhere else:

> Why do we need theory if you have incontrovertible evidence of efficacy? My argument is that if you don’t have both you don’t just have half or whatever of what you should have, but that you have nothing. We all recognize that theory without evidence to support it leads to no conclusions. The reverse is true as well. (Cartwright, 2009, p. 205)

Ideal RCTs are not the only methodology for arriving at conclusions; other methods that work deductively can *clinch* them. Econometric methods, Galilean experiments, probabilistic/Granger causality, derivation from established theory and tracing the causal process can also do the work: “[T]here is no *a priori* reason to favor a method that is rigorous part of the way and very iffy thereafter over one that reverses the order or one that is less rigorous but fairly well reasoned throughout” (Cartwright, 2013, p. 31). Cartwright also suggests that sometimes RCTs are not needed and that simple observation can do the job without putting the population at risk (Deaton & Cartwright, 2018, p. 7). As Judea Pearl posits on a commentary to Deaton and Cartwright’s work: “In addition, considering the practical difficulties of conducting an ideal RCT, observational studies have a definite advantage: they interrogate populations at their natural habitats, not in artificial environments choreographed by experimental protocols” (Pearl, 2018, p. 60).

Considering the limitations of RCTs, what can we use them for? Cartwright provides some answers to the question: they may yield a type of falsification test to refute a theoretical proposition, or they may confirm the prediction of theory (not the theory itself). They are also able to show that a specific treatment does, in fact, work in a specific setting (Deaton & Cartwright, 2018, p. 13).

In evidence-based public policy, the effort should focus on the effectiveness of the treatment, not only its efficacy. The question, therefore, is how to warrant that prediction. As Cartwright posits, a “warrant requires a good argument” (Cartwright, 2012, p. 15), an argument that has to be sound and valid, with premises that can be trusted, and a conclusion that is really implied by those premises. Good arguments lead us to robust conclusions. Accordingly, to check for effectiveness, policymakers need a good argument based on facts about causal principles and not just simply statistical

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

associations. The type of causal principles that experts need to find in social sciences are *ceteris paribus*: they require that all conditions remain the same. These types of principles can be deterministic or probabilistic.

Finding the causal principles of a policy requires that we look into the support factors that make the policy work. Policies need contributing structures set in place for them to work – they do not work by themselves. They are part of a team of structures that belong to a setting. To predict effectiveness, experts need to understand those support factors. If that were the case, the study would have external validity, since the treatment obtains the same result as it did in the initial study. As María Jiménez-Buedo explained, while internal validity deals with how reliable causal inferences are, external validity tackles the possibility of generalizing them to events beyond the experimental setting (Jiménez-Buedo, 2011, p. 271). Experts in the industry suggest that we can get such external validity when the same treatment is applied to a population that is sufficiently similar to the one from the initial study. Cartwright argues that this perspective is too vague since it does not help with the effectiveness of a policy: applying the same treatment to obtain the same result in a different setting leaves out the actual causal principles, and the requirement for *similarity* is too demanding and wasteful (Cartwright, 2012, pp. 46–49). The external/internal validity dichotomy is not exempt from criticism. Jiménez-Buedo suggests that due to certain conceptual problems, this distinction is prone to ambiguous interpretations. She wonders about the notion of validity and what it refers to: Is it the experiment, the type of experiment, the experimental results, the data obtained from experiments, etc. (Jiménez-Buedo, 2011, p. 274)? Cartwright warned us about the danger of substituting deliberate thinking regarding causal factors with just external validity: "Only by thinking in terms of causal roles and support factors can you begin to see what evidence you need if you are going to bet that the policy will work here. You cannot avoid thinking like that. The notions of external validity and similarity are no substitutes" (Cartwright & Hardie, 2012, p. 49). The notions of external and internal validity are not practical for extracting accurate theoretical interpretations from observed events (Jiménez-Buedo, 2011, p. 279).

Accordingly, Cartwright summarizes that the argument to verify the effectiveness of a policy should go as follows: the policy worked in one setting, the policy can play the same causal role in a different place and the support factors that helped the policy to play a positive causal role are shared by at least some individuals in both settings (Cartwright & Hardie, 2012, p. 54). Those premises would yield the prediction that the policy will work on the other settings.

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

There are several elements to consider in these situations:

- Causal principles can change once a policy is in place after an RCT has confirmed its efficacy.
- The deployment of a policy may affect the causal structure of a setting to the point that the policy itself no longer obtains the desired effect.
- The same policy may carry positive consequences for a group of individuals and negative consequences for another group (Cartwright & Hardie, 2012, p. 45).
- RCTs are not exempt from producing moral dilemmas: randomization may expose people to unnecessary issues, especially in medicine or public policy matters, which may dramatically affect the lives of individuals (Deaton & Cartwright, 2018, p. 7).

There are instances where if we apply a higher level of abstraction we can find causal principles that are common in different settings. If a factor is able to play the same causal role in different situations, that factor is acting at a more abstract level. Since support factors and causal principles can verify whether a policy may work in a different setting, and policies work in tandem with other variables (*ceteris paribus*); if those variables are not part of the setting, their absence will make the policy ineffective. This idea of causation is well explained by E.J. Mullen:

> What is meant by causal principles and support factors? Cartwright and Hardie borrow a currently popular philosophical view of causation attributed to J.L. Mackie which proposes that causes are "at a minimum INUS conditions, that is, 'Insufficient but Necessary parts of a condition which is itself Unnecessary but Sufficient' for their effects" … This view of causation proposes that typically causes come in clusters of variables (rather than a single causal agent) and that there are conceivably many such clusters for any given effect. Applying this logic to social intervention it suggests that any given intervention must be accompanied by other variables in the cluster to have the intended effect, but that aside from the given intervention other possible clusters of variables which do not include this intervention could result in the same effect. (Mullen, 2016, p. 320)

In order to find support factors, we need to do what Cartwright called a "horizontal search" (Cartwright, 2012, p. 91). If an RCT provides a positive result about a specific policy, its success would be related to a mix of factors that help the treatment to be effective. The other elements needed to verify that a particular policy may work in a different setting are, as mentioned before, causal principles. Those causes can have different levels of abstraction. A *vertical search* more or less would go up and down through those levels. Vertical searches should be able to determine the appropriate level of abstraction to retrieve

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

common explanatory factors that may intervene in how a policy may yield a positive outcome (Cartwright, 2011b, p. 23).

Experts usually already have an idea of how the intervention may cause the desired outcomes before the RCT. It is not just trial and error, since those experts rely on previous knowledge to bet that the intervention will work, speculating on the causal principles and the support factors needed for it to be successful. Science will provide for you that type of background knowledge. In social sciences, one needs to find and understand what happens and the causal roles of the intervention. That is what vertical and horizontal research will give you (Cartwright & Hardie, 2012, p. 126). Beyond using RCTs, one must think and deliberate: "Deliberation is not second best … To deliberate in order to exercise discretion requires a rich list of intellectual and practical virtues that cannot be reduced to the virtue of conformity. Thus, the orthodoxy not only discourages deliberation as unnecessary, since the rules are superior but selects in favor of operatives who cannot deliberate" (Cartwright & Hardie, 2012, pp. 158–159). Deliberation may increase the external validity of an experiment, opening up the possibility of answering the question: "Can we use experimental knowledge to understand what goes on in the 'real world'?" (Guala, 2012, p. 612). When researchers conduct an RCT seeking external validity, they should immerse themselves in a process of deliberation to provide inferences that will yield that validity. Deliberation should help with what Francesco Guala called the "justification" of the inference process (Guala, 2012, p. 612).

If mistakes are made while thinking, the policies will not be effective, even if RCTs are used. Creating rules to avoid thinking eliminates a necessary part of the process of discovering the effectiveness of policies.

Nudge theory

> A nudge, as we will use the term, is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates. Putting the fruit at eye level counts as a nudge. Banning junk food does not. (Thaler & Sunstein, 2009, p. 6)

We have examined Cartwright's analysis of evidence-based policies and RCTs. The following pages will be devoted to nudges as evidence-based policy interventions with the objective of analyzing their epistemological status using Cartwright's philosophical approach. These pages will also show how the consideration of expressive rationality can make nudges more robust while attending to ethical considerations related to people's autonomy. Accordingly,

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

this section starts with a short history of nudge theory and the psychological framework behind its use, to help us examine how nudges conceive human rationality and how this conception may be problematic. It will conclude by showing how the consideration of expressive rationality and the application of Cartwright's approach by weighing the causal principles and support factors of nudges within a deliberative process may allow policymakers to obtain more effective results while respecting the autonomy of those who are nudged.

Most nudges rely on an instrumental view of rationality. Holding a narrow view of rationality within a means–end schema, nudges only focus on how treatments can positively affect the outcome of a decision and if behavior satisfies the goals of the individual. Accordingly, the most rational agent will select the appropriate means to maximize their utility function. Nudges focus only on the instrumental part of rationality and consider irrational the behavior that will not yield maximization. A nudge is not only a tool to help the instrumental rationality of individuals to become more effective, but also an intervention to increase social welfare by changing people's behavior (e.g., increase organ donation). Most policymakers that use nudges tend to neglect the reasons why individuals behave in the way they do or why they modify their decisions. To verify whether those nudges will work during a period of time in the same or a different setting, policymakers test them, as mentioned above, mostly using RCTs (Einfeld, 2019, p. 509). To provide a measure of the scale of the use of RCTs by nudge units, a recent analysis on nudges conducted by Stefano DellaVigna and Elizabeth Linos (2020) used data from the two biggest nudge units in the USA: it included 126 RCTs and 23 million people. Anneliese Arno and Steve Thomas, for example, conducted a systematic review of current nudge papers to determine whether these strategies "are successful in changing adults' dietary choices for healthier ones" (Arno & Thomas, 2016, p. 1). For the review, they used 37 papers, 31 of which (74%) reported RCTs.

Nudges have the possibility to fail or backfire, and some authors argue that they may even pose a limitation to the autonomy of the decision, since they sometimes manipulate the choice environment without the awareness of those they are trying to nudge (McCrudden & King, 2015).

Ever since Herbert Simon coined and developed the notion of bounded rationality (Simon, 1957) to describe the processes of how people decide, numerous authors have approached social sciences from that perspective, considering that our rationality is bounded by cognitive limitations and by the complex structure of the environment. While standard economics presumed an economic agent with unlimited cognitive skills, infinite memory and capable of optimizing their decisions, Simon sustained that a robust empirical

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

foundation of economics contradicted that view. When Simon introduced psychology as part of the theoretical framework for understanding specific phenomena in social sciences, providing a more realistic description of economic decisions, a new type of social agent appeared: one with bounded rationality. Simon claimed that people *satisfice* instead of optimize. The rationality of our decisions, therefore, had to be understood within the boundaries imposed by our rationality and the structure of the environment.

The idea that our rationality was limited was also later developed by other authors. The work of Amos Tversky and Daniel Kahneman in heuristics and biases (Kahneman & Tversky, 1979; Tversky & Kahneman, 1974, 1981) showed that our rationality systematically errs, concluding that people "suffer" from a set of biases that makes them "irrational." They claimed that the systematicity of these errors allows for their prediction. Following that view, some public policymakers suggest that certain interventions can be proposed in a way that might respond to those failures without limiting the freedom of the agent. That biases are systematic, epistemologically speaking, indicates that we can scientifically operate with them and predict their occurrence. Social scientists believe that RCTs can be used to predict the effectiveness of their interventions, but as mentioned before, no evidence supports the possibility of that prediction. I will return to this issue below.

Richard Thaler and Cass Sunstein, following the work of Tversky and Kahneman (1974), proposed a solution to people's irrationality without disturbing their freedom to choose (Thaler & Sunstein, 2009): nudges. Tversky and Kahneman are not themselves nudge theorists; rather, nudge theory rests on the research work on biases and heuristics conducted by these authors. They recognized the systematicity and predictability of people's rational errors. Consequently, since people's decisions are inserted in a choice environment and that it is costly and sometimes impossible to educate them, nudge theorists maintain that the choice architect can organize the environment of the decision in a way that would take into consideration the biases of individuals to increase the efficiency (instrumental) of public policy interventions regarding behavioral change.

Considering that nudges are able to promote change while respecting freedom of choice, different nudge units have been created throughout the world by governments and institutions captivated by their success and promises of efficiency: the Behavioural Insights Team (BIT), the Ontario Behavioural Insights Unit (OBIU), the Behavioural Insights Network Netherlands, MineduLab in Peru, the Behavioural Economics Team in Australia, etc. The BIT emphasizes above all the use of RCTs for testing purposes. David Halpern, chief executive at BIT, promotes their use:

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

> Although we know there is a set of factors that influence behavior, we don't know for certain which will apply in a particular context. Therefore, BIT has promoted a "Test, Learn, Adapt" approach to government, based around the use of randomized controlled trials (RCTs). RCTs have a reputation in government for being expensive, difficult to implement, and slow to give results. BIT has set about showing that they can be cheap and feasible, and can give quick feedback to improve policy making. (Halpern, 2013)

A nudge can be a change in the location of a salad bar in a university cafeteria, the placement of a signature on a tax form or a default option to save 6% of your salary for retirement. Nudges can be successful and, if they fail, proponents argue that problems are limited, since they should be easily avoided by individuals. Cass Sunstein indicates that even if nudges are not effective, they should be kept if they improve people's welfare: "A largely ineffective nudge may have positive welfare effects; an effective nudge might turn out to reduce welfare. A strong reason for nudges, as distinguished from more aggressive tools, is that they preserve freedom of choice and thus allow people to go their own way" (Sunstein, 2017, p. 22).

Expressive rationality and nudge theory

In economics and social sciences, rationality is commonly defined instrumentally (Nozick, 1993, p. 133): it helps individuals to choose the most efficient means to get to their desired goals. Nudges, from a philosophical perspective, look at decisions through that instrumental lens, aiming at altering the outcomes of behaviors by manipulating the choice environment. Other authors (Boudon, 1998; Álvarez, 2002; Echeverría & Álvarez, 2008; Bouwmeester, 2017) claim that rationality is not only instrumental, but also expressive: people have reasons to act in a specific way, sometimes regardless of the consequences. Often people want to express their subjective selves with their actions.

In order for nudges to be more efficient, besides considering the outcome of decisions, they should try to understand rationality comprehensively, by examining its bounded (Bouwmeester, 2017, p. 3), social and expressive components (Hortal, 2019). Conceiving rationality only within an instrumental frame would neglect essential elements of rational decisions, impeding a full understanding of the behavior. Accordingly, nudges should include the instrumental and bounded aspects of rational behavior (as they are currently doing), understanding at the same time the expressive and the social rationality of the causal factors of group behavior (Bicchieri, 2017, p. 48).

What is important for our instrumental rationality is that we are aware of the causes and effects of our behavior, so we can decide efficiently. According to

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

this frame, rationality's spectrum would have two poles: more rational, if you choose the means that yield the desired outcomes; and less rational, if you choose something less desirable. Our instrumental rationality focuses on how to be efficient in order to arrive at our goals, but it says nothing about the goals themselves. Nudge theory, examining rationality within just an instrumental frame, views people as fundamentally irrational. Nudge theorists complement this view by using the two-system cognitive model proposed by Kahneman (2011) to understand decisions, where system 1 is automatic and intuitive and system 2 is deliberative. That distinction is not exempt from criticism. Gerd Gigerenzer, for example, claims that this division is rather problematic due to its vagueness:

> It makes it possible to explain everything after the fact but not to deduce any interesting novel prediction. Usually, science progresses from vague dichotomies to precise models; the two-systems story is the only case I know of where it went the other way. Behavioral economists have reduced existing mathematical models of heuristic and statistical inference to two black boxes. (Gigerenzer, 2015, p. 379)

When we behave, we are not only trying to achieve a desired goal (e.g., losing weight, saving more money) – we also have reasons for doing what we are doing. Those reasons can be social or expressive. Social reasons are anchored to intersubjective elements, while expressive reasons are subjective. Behavior is complex and is grounded on the three different components of expressive, social and instrumentally rationality. For example, if we vote in a general election, our instrumental rationality may have the illusion that our vote matters and our action will have consequences (Opp, 2015, p. 191). We also have good reasons to vote: our social rationality may see it as a norm, and our expressive rationality may enjoy the act of voting: "Expressive rationality refers to what we value subjectively or as a single person. Expressive arguments can, for instance, explain the rationality of what we do during a holiday, how we dress, what music we listen to, etc. Acting on personal impulses, interests, or motives can make our actions rational from a subjective perspective" (Bouwmeester, 2017, p. 8). Accordingly, we can explain actions within different rational frames, but we should assume a comprehensive model of rationality. In the previous example, the comprehensive approach would conclude that voting can be instrumentally irrational while expressively and socially rational.

The reductionist instrumental approach disregards part of the reasons behind human behavior, explaining everything using a vague two-system theory. This view causes the theory of nudges to rest on an incomplete theory of human rationality. The comprehensive theory, which includes expressive and social rationality as part of the theoretical frame, can make

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

nudges more efficient, and since those expressive aspects of rationality act as causal principles of behavior, an understanding of these aspects can provide much-needed evidence for the use of nudges in settings that go beyond the initial one targeted by the RCT. Cristina Bicchieri (2017) and other researchers (Bicchieri & Dimant, 2019) are already working on the social aspects of nudges in a way that regards our social rationality as part of the reasons behind behavior without stigmatizing these social aspects as irrational. Some authors, for example, are using these social aspects to test interventions related to energy consumption (Brandon *et al.*, 2019) or vaccination (Korn *et al.*, 2018).

There is a major issue when including expressive rationality in nudge theory: its subjective aspect. While nudges consider the instrumental and bounded aspects of rationality, they do so by attending to the systematic biases of human rationality. They rest on universal and systematic aspects of human cognition. When social nudges are deployed, they are intervening in a social context where, more or less, experts can obtain information about the social norms of the setting. It is difficult to gather data or information regarding expressive elements that can be used by policymakers when researching possible nudges.

Nudges should consider expressive rationality in two different ways: to respect the subjective autonomy of individuals and to make interventions more efficient. Our expressive rationality is molded culturally, so it is possible that some expressive components are shared by large groups of individuals in the same setting. If that is the case, research can be conducted to include these expressive components in behavioral interventions.

For example, if the purpose of a nudge is to change people's behavior so that they can eat healthier, the most effective nudge would understand:

- The bounded rationality of individuals in their instrumental actions
- Their social rationality and how social norms play a role in how we eat
- Their expressive rationality and how tastes, likes and dislikes affect what we eat and how we eat it

To this end, that intervention would not only change the environment to position a salad in a cafeteria in a way that may increase its consumption – it would also try to make salad consumption a social norm while making it tastier (or by introducing other expressive elements such as soccer teams for salad names, etc.).

The idea, therefore, is to make sure that nudges are aware of the diverse components of our rationality, and that these components are included in the search for the causes of our behavior and for the causes of why policies work in a particular setting. Accordingly, understanding those other types of rationality should be a must for nudge theorists. Our decisions are a

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

complex system in which those different rationalities interact. Our beliefs, thoughts and tastes – that is, what Pierre Bourdieu called *habitus* (Bourdieu, 1985) – are essential causes of behavior. The notion of *habitus* represents a group of tendencies that lead our behavior and form our identity, and it is the result of social externalities related to class and other circumstances. Nudges should consider this habitus – the social and expressive components of behavior – to understand the differences in the effects of interventions.

Concluding remarks: expressive rationality, Cartwright's causal principles and nudge theory

In the first section of this paper, we reviewed Cartwright's philosophy and examined how she argued that evidence-based policies should try to understand causal principles and support factors besides using RCTs to provide more robust evidence for their effectiveness. To this end, social scientists should search horizontally and vertically to find those principles and factors. Once they are found, they can have a better and more comprehensive understanding of the policies and the mechanisms that make them work.

Nudges, using behavioral insights focusing on cognitive biases, change people's behavior by manipulating the choice environment. As evidence-based policies, they heavily rely on RCTs. As Jiménez-Buedo argues, an RCT does not guarantee the possibility of forming a robust inference regarding a causal relationship (Jiménez-Buedo, 2011, p. 274). Consequently, I claim that nudges can be improved if we apply Cartwright's epistemological recommendations mentioned above. Another improvement can be achieved by examining people's expressive rationality, which acts as a causal principle regarding their behavior. Understanding the reasons as to why people behave will increase the effectiveness of nudges and provide understanding of how they can be deployed while respecting people's subjective autonomy. This type of understanding, using Cartwright's words, will allow experts to "climb up the ladder of abstraction," increasing trust in the specific nudge intervention. Our expressive rationality acts as a causal principle of our behavior. Understanding the reasons as to why people behave will increase the chances of collecting more evidence regarding the effectiveness of nudges. For example, a study conducted to increase influenza vaccination rates among employees (Milkman *et al.*, 2011) showed that mailing people reminders listing the time and location of the appointment with a prompt instructing them to write down the date increased the vaccination rate by 1.5% (which was statistically insignificant). If, instead of just the date, the prompt also asked the recipient to write down the time of the appointment, there was an increase in the vaccination rate of 4.2%. Was the change related to just

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

adding the actual time of the appointment? It may be the case, for example, that what really caused the behavioral change was not related to writing down the time of the appointment, but the possibility of writing something about the appointment so that individuals can spend a few more seconds thinking about the appointment. If we understand what caused the increase, we can implement it in other settings.

Text messages, for example, have been used successfully to increase loan repayments in Uganda (Cadena & Schoar, 2011), to improve the collection of overdue fine payments (Behavioral Insights Team, 2012) and to increase commitments to savings (Karlan *et al.*, 2016). The same intervention (text message) did not work to increase the number of future students completing FAFSA, a form for obtaining financial aid for college (Bird *et al.*, 2019). Text messages also did not work to improve academic performance (Oreopoulos & Petronijevic, 2019). Why do text messages work in certain scenarios and not in others? Understanding what causes the above-mentioned behavioral changes on a more abstract level than the medium of the intervention can improve the effectiveness of policies. People's expressive (and social) rationality is involved in their decisions – understanding their different levels of influence in specific decisions will improve the effectiveness of nudges.

Reminders are also a good example of how nudges work. Trying to increase adherence to medications prescribed after heart attacks, Volpp *et al.* (2017) tested the effect of electronic reminders (nudge) under the assumption that they would increase such adherence. Contrary to their initial hypothesis, an RCT conducted with more than 1500 patients showed that those nudges caused no improvement. These authors thought that since similar nudges had worked in other settings, they would work on their own patients: "Adherence to medications prescribed after acute myocardial infarction (AMI) is low. Wireless technology and behavioral economic approaches have shown promise in improving health behaviors" (Volpp *et al.*, 2017, p. 1093). As they mentioned, similar approaches "have shown promise," but this promise is not a warrant for effectiveness. If people's adherence to medication does not improve after specific policies are in place, more RCTs with different nudges are not going to fix the situation without a comprehensive understanding of the reasons behind people's behavior. The solution cannot simply consist of trialing new interventions without trying to understand the expressive rationality of individuals. For example, the research mentioned above claimed that behavioral economics could offer "promise in improving motivation for desirable but difficult activities, such as weight loss, exercise, or smoking cessation by harnessing pervasive patterns of irrational behavior" (Volpp *et al.*, 2017, p. 1094). Treating human rationality as a black box and establishing nudges by testing their outcomes while disregarding expressive

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

or social reasons for behavior, all while defining non-normative behavior as irrational, can make nudges ineffective. Throughout their research article, not a single word is mentioned on the possible reasons for the lack of adherence. However, the authors did observe that due to the negative results of the intervention, different approaches need to be considered (Volpp *et al.*, 2017, p. 1099).

After facing negative results in the outcomes of policies based on nudges, authors tend to request more research. Using Cartwright's perspective, what nudges need is a better understanding of expressive and social rationality. They both act as causal principles and as determinants in order for policies to work. Nudges should also pay attention to the support factors involved in policies, not simply testing policies using RCTs. They should use vertical searches to determine the expressive and social reasons underlying people's behavior. They should also use horizontal searches to discover the support factors that might help to explain why some policies fail while others succeed. Even if a causal link has been found in an experimental setting, it will be difficult to extrapolate the finding to the real world, since we may not be aware of all of the possible factors contributing to the effect (Jimenez-Buedo & Miller, 2010). By understanding the expressive rationality of individuals in their research, policymakers can also ensure that the autonomy of the individual in their decision-making process is respected. A nudge that positively alters the choices people make by tackling the reasons behind their actions is a good method for ensuring respect of their autonomy. Behavioral changes that bypass people's awareness can pose serious moral and political problems.

Consequently, in conjunction with RCTs, policymakers should think and deliberate about the causes of behavior, its changes and the causal mechanisms of policies. Cass Sunstein sometimes has argued for a similar idea: "[I]f a nudge is based on a plausible but inaccurate understanding of behavior, and of the kinds of things to which people respond, it might have no impact" (Sunstein, 2017, p. 20). I regard this task as being rather difficult if our expressive rationality is not considered as one of the causal principles involved in our behavior and, consequently, as a subject of study for policymakers.

In future research, it would be interesting to define how experts can "climb the ladder of abstraction" to find the causes of behavioral change and how they relate to specific nudges considering the different aspects of rationality. In the vertical search for causes, the application of Pierre Bourdieu's sociology, particularly his work on habitus and social fields, may shed some light on how nudges may affect people differently due to inequality and other social elements that act as causal principles. With habitus, Bourdieu refers to tastes, beliefs, interests and our understanding of the world. These elements, as integral

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

parts of the non-instrumental aspects of our rationality, should be included in any research regarding nudges.

References

Álvarez, J. F. (2002), 'El tejido de la racionalidad acotada y expresiva', *Manuscrito*, **XXV**, 11–29.

Arno, A., S. Thomas (2016), 'The efficacy of nudge theory strategies in influencing adult dietary behaviour: a systematic review and meta-analysis', *BMC Public Health*, **16**, 676. doi: 10.1186/s12889-016-3272-x

Behavioral Insights Team (2012), Test, Learn, Adapt. Developing Public Policy with Randomised Controlled Trials.

Bicchieri, C. (2017), *Norms in the wild: how to diagnose, measure, and change social norms*. Oxford University Press. doi:10.1093/acprof:oso/9780190622046.001.0001

Bicchieri, C., E. Dimant (2019), *Nudging with Care: The Risks and Benefits of Social Information* (No. ID 3319088). Social Science Research Network, Rochester, NY.

Bird, K., B. Castleman, J. Denning, J. Goodman, C. Lamberton, K. O. Rosinger (2019), *Nudging at Scale: Experimental Evidence from FAFSA Completion Campaigns* National Bureau of Economic Research, Cambridge, MA. doi:10.3386/w26158

Boudon, R. (1998), 'Limitations of Rational Choice Theory', *The American Journal of Sociology*, **104**, 817–28.

Bourdieu, P. (1985), 'Distinction: A social critique of the judgement of taste', *MLN*, **100**, 1133. doi: 10.2307/2905454

Bouwmeester, O. (2017), *The social construction of rationality: policy debates and the power of good reasons*. Routledge, Abingdon, Oxon; New York, NY: Routledge, 2017. |. doi:10.4324/ 9781315724379

Brandon, A., J. A. List, R. D. Metcalfe, M. K. Price, F. Rundhammer (2019), 'Testing for crowd out in social nudges: Evidence from a natural field experiment in the market for electricity', *Proc Natl Acad Sci USA*, **116**, 5293–5298. doi:10.1073/pnas.1802874115

Cadena, X., A. Schoar (2011), *Remembering to pay? reminders vs. financial incentives for loan payments* National Bureau of Economic Research, Cambridge, MA. doi:10.3386/w17020

Cartwright, N. (2009), *What is this thing called "efficacy"?. Philosophy of the social sciences : philosophical theory and scientific practice* Cambridge University Press. 185–206.

Cartwright, N. (2011a), 'A philosopher's view of the long road from RCTs to effectiveness', *Lancet*, **377**, 1400–1401. doi:10.1016/s0140-6736(11)60563-1

Cartwright, N. (2011b), Evidence, External Validity, and explanatory relevance, in: G. J. Morgan (Ed.), *Philosophy of Science Matters: The Philosophy of Peter Achinstein* Oxford University Press, pp. 15–28.

Cartwright, N. (2012), *Rcts, evidence, and predicting policy effectiveness, Oxford Handbooks Online* Oxford University Press. doi:10.1093/oxfordhb/9780195392753.013.0013

Cartwright, N. (2013), Evidence : for policy and wheresoever rigor is a must., Order Project Discussion Paper Series. London School of Economics and Political Science (LSE)., London.

Cartwright, N. (2018), 'What evidence should guidelines take note of? *J. Eval. Clin. Pract*. **24**, 1139–1144. doi:10.1111/jep.12959

Cartwright, N., J. Hardie (2012), *Evidence-Based Policy: A Practical Guide to Doing It Better* Oxford University Press. doi:10.1093/acprof:osobl/9780199841608.001.0001

Deaton, A., N. Cartwright (2018), 'Understanding and misunderstanding randomized controlled trials', *Soc. Sci. Med*. **210**, 2–21. doi:10.1016/j.socscimed.2017.12.005

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

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

Echeverría, J., J. F. Álvarez (2008), Bounded Rationality in Social Sciences, in: E. Agazzi (Ed.), *Epistemology and the Social. Rodopi*, pp. 173–91.

Einfeld, C. (2019), 'Nudge and evidence based policy: fertile ground', *Evid. Policy*, **15**, 509–524. doi: 10.1332/174426418X15314036559759

Gigerenzer, G. (2015), 'On the supposed evidence for libertarian paternalism', *Rev. Philos. Psychol.* **6**, 361–383. doi:10.1007/s13164-015-0248-1

Guala, F. (2012), *Experimentation in Economics, in: Philosophy of Economics* Elsevier, 597–640. doi:10.1016/B978-0-444-51676-3.50021-X

Halpern, D. (2013), 'Applying psychology to public policy', *APS Observer*, **27**.

Hortal, A. (2019), 'Nudging and educating: bounded axiological rationality in behavioral insights. Behav', *Public Policy* 1–24. doi:10.1017/bpp.2019.2

Jiménez-Buedo, M. (2011), 'Conceptual tools for assessing experiments: some well-entrenched confusions regarding the internal/external validity distinction', *Journal of Economic Methodology*, **18**, 271–282. doi:10.1080/1350178X.2011.611027

Jimenez-Buedo, M., L. M. Miller (2010), 'Why a trade-off? The relationship between the external and internal validity of experiments', *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia*, **25**, 301–321.

Kahan, D. M. (2017), 'The expressive rationality of inaccurate perceptions', *Behav. Brain Sci.* **40**, e6. doi:10.1017/S0140525X15002332

Kahneman, D. (2011), *Thinking, Fast and Slow* Macmillan.

Kahneman, D., A. Tversky (1979), 'Prospect Theory: An Analysis of Decision under Risk', *Econometrica*, **47**, 263. doi:10.2307/1914185

Karlan, D., M. McConnell, S. Mullainathan, J. Zinman (2016), 'Getting to the top of mind: how reminders increase saving', *Management Science*, **62**, 3393–3411. doi:10.1287/mnsc.2015.2296

Korn, L., C. Betsch, R. Böhm, N. W. Meier (2018), 'Social nudging: The effect of social feedback interventions on vaccine uptake', *Health Psychol.* **37**, 1045–1054. doi:10.1037/hea0000668

McCrudden, C., J. King (2015), 'The dark side of nudging: the ethics, political economy, and law of libertarian paternalism', *Choice Architecture in Democracies, Exploring the Legitimacy of Nudging* (Oxford/Baden-Baden: Hart and Nomos, 2015), Forthcoming.

Milkman, K. L., J. Beshears, J. J. Choi, D. Laibson, B. C. Madrian (2011), 'Using implementation intentions prompts to enhance influenza vaccination rates', *Proc Natl Acad Sci USA*, **108**, 10415–10420. doi:10.1073/pnas.1103170108

Mullen, E. J. (2016), 'Reconsidering the 'idea'of evidence in evidence-based policy and practice', *European journal of social work*, **19**, 310–335.

Nozick, R. (1993), *The Nature of Rationality* Princeton University Press, Princeton, N.J.

Oliver, A. (2019), *Reciprocity and the art of behavioural public policy* Cambridge University Press. doi:10.1017/9781108647755

Opp, K. D. (2015), 'Instrumental, Axiological Rationality and the Explanation of Norms. Cherkaoui's (and Boudon's) Critique of Rational Choice Theory and Its Ability to Explain Norm Emergence', *Theories and Social Mechanisms. Essays in Honor of Mohamed Cherkaoui*, **1**, 183–206.

Oreopoulos, P., U. Petronijevic (2019), *The Remarkable Unresponsiveness of College Students to Nudging And What We Can Learn from It* National Bureau of Economic Research, Cambridge, MA. doi:10.3386/w26059

Pearl, J. (2018), 'Challenging the hegemony of randomized controlled trials: A commentary on Deaton and Cartwright', *Soc. Sci. Med.* **210**, 60–62. doi:10.1016/j.socscimed.2018.04.024

Simon, H. (1957), *Models of man: social and rational; mathematical essays on rational human behavior in society setting* Wiley.

Sunstein, C. R. (2017), 'Nudges that fail. Behav', *Public Policy*, **1**, 4–25. doi:10.1017/bpp.2016.3

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press

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

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

Tversky, A., D. Kahneman (1981), 'The framing of decisions and the psychology of choice', *Science*, **211**, 453–458. doi:10.1126/science.7455683

Volpp, K. G., A. B. Troxel, S. J. Mehta, L. Norton, J. Zhu, R. Lim, W. Wang, N. Marcus, C. Terwiesch, K. Caldarella, T. Levin, M. Relish, N. Negin, A. Smith-McLallen, R. Snyder, C. M. Spettell, B. Drachman, D. Kolansky, D. A. Asch (2017), 'Effect of electronic reminders, financial incentives, and social support on outcomes after myocardial infarction: the heartstrong randomized clinical trial', *JAMA Intern. Med.* **177**, 1093–1101. doi:10.1001/jamainternmed.2017.2449

https://doi.org/10.1017/bpp.2020.55 Published online by Cambridge University Press