2022

The Law and Economics of Behavioral Regulation

Avishalom Tor

Avishalom Tor*

The Law and Economics of Behavioral Regulation

https://doi.org/10.1515/rle-2021-0081
Published online July 13, 2022

Abstract: This article examines the law and economics of behavioral regulation ("nudging"), which governments and organizations increasingly use to substitute for and complement traditional instruments. To advance its welfare-based assessment, Section 1 examines alternative nudging definitions and Section 2 considers competing nudges taxonomies. Section 3 describes the benefits of nudges and their regulatory appeal, while Section 4 considers their myriad costs—most notably the private costs they generate for their targets and other market participants. Section 5 then illustrates the assessment of public and private welfare nudges using cost-benefit analysis, cost-effectiveness analysis, and rationality-effects analysis.

Keywords: nudge, rationality, bounded rationality, regulation, cost–benefit analysis, cost-effectiveness analysis, rationality-effects analysis

JEL Classification: D61, D90, D91, H40, K29

1 Introduction: The New Behavioral Regulatory State

Behavioral regulation—commonly referred to as nudging—is on the rise (Mathis and Tor 2016; Jones et al. 2013; Oliver 2017; Tor 2016, 2019, 2020a). For over a decade, governments and other organizations have been increasingly turning to these "soft" behavioral interventions to achieve their policy goals. By now, nudges have already been implemented in nearly every major policy domain

Professor of Law and Director, Notre Dame Research Program on Law Market Behavior (ND LAMB), Notre Dame Law School; Individual Fellow, Israel Institute for Advanced Studies. This project benefited from the generous support of Notre Dame Law School and the Israel Institute for Advanced Studies. Isabella Wilcox provided excellent research assistance.

***Corresponding author: Avishalom Tor**, University of Notre Dame, Notre Dame, USA,
E-mail: ator@nd.edu

that concerns individual behavior, from health, safety, education, and finance through environmental protection and tax compliance, to public service delivery and more. In all of these areas, regulators aim to promote private or public welfare by shaping the behavior of the people their policies target (Tor 2019). Reports by the European Commission (2016) and the Organisation for Economic Co-operation and Development (2017) detail over 100 case studies of behaviorally-informed interventions in Europe, North America, and elsewhere, while the U.K.-based Behavioural Insights Team (2019)—the most active among the various international organizations in this field—recently reported having run more than 780 projects in dozens of countries since 2010.

Nudging draws on the evidence and methods of behavioral science to inform policy design (Madrian 2014). This approach focuses on the novel legal prescriptions suggested by evidence of systematic differences between real human behavior and the hypothetical rationality that standard economic models assume (Tor 2016). In particular, real people possess limited cognitive resources and are affected by motivation and emotion—in short, they are only "boundedly rational". Although they sometimes engage in formal, effortful, and time-consuming judgment and decision making, to function successfully in a complex world, individuals commonly employ mental and emotional heuristics to make intuitive judgments under uncertainty and rely on situational cues to guide their choices. These heuristic judgments and cue-dependent choices are highly adaptive and often useful, but they also lead decision makers systematically and predictably to deviate from the normative standards of strict rationality (Tor 2008).

While understanding these differences is necessary even for traditional regulation to achieve its goals (Madrian 2014), it is particularly critical for nudging, which is distinguished by its reliance on significantly behavioral effects that the hypothetical rational actor would have found largely irrelevant (Bernheim and Taubinsky 2018; Tor 2016). In this respect, therefore, behavioral policies differ from traditional instruments that shape people's behavior by changing their constraints (as mandates or bans do) or economic incentives (as in the case of taxes or subsidies), or by providing mere information that may otherwise be unavailable or costly to obtain (like traditional disclosure requirements) (Allcott and Sunstein 2015; Thaler and Sunstein 2008; Tor 2021a). Instead, nudges use non-coercive behavioral tools (Thaler and Sunstein 2008; Tor 2019) to guide people's behavior, through more effective or persuasive information presentation, the framing of the alternative choices available, the selection of defaults, the shaping or communication of social norms or other social information, and more (Sunstein 2016).

By now, behavioral interventions are used extensively, as both substitutes for and complements to traditional regulation. Around the world, national responses

to the coronavirus pandemic have vividly illustrated this trend, as many governments employed behavioral campaigns as substitutes for using traditional instruments, such as mandates or financial incentives, in an effort to encourage widespread vaccination; at the same time, public authorities also used nudges to complement and increase the efficacy of quarantine or masking mandates, given the exceedingly high costs and limited efficacy of their traditional public enforcement (Teichman and Underhill 2021).

Regulators are attracted to nudging for both principled and pragmatic reasons. For one, insofar as behavioral regulation is based on a more realistic view of the regulated than that offered by the assumptions of traditional economic models, public policy makers may find nudging intuitively more compelling. Because it employs non-coercive instruments, moreover, nudging appeals to governments and other institutions in democratic nations that value citizen freedom and autonomy (Sunstein 2015). Related, political actors who believe that the public also finds nudges more acceptable than "harder", traditional regulation will tend to prefer the former to the latter when possible (Sunstein and Reisch 2019). Since they draw on a host of psychological, emotional, and other behavioral processes, moreover, nudges make highly versatile policy tools that—at least in principle—can be designed and fine-tuned to address a broad range of policy challenges (Sunstein 2016). Finally, and perhaps most importantly for its widespread adoption, behavioral regulation is viewed as a low-cost approach that offers governments an opportunity to advance important policy goals while imposing only a limited burden on strained public budgets (Sibony and Alemanno 2015; Sunstein and Reisch 2019).

As attractive as nudges appear to regulators and scholars, it is their welfare effects that truly matter from a law and economics perspective. Behavioral regulation can increase both individual and aggregate social welfare. Crucially, though, despite operating through "soft" behavioral means, nudges—like other regulatory instruments—are rarely cost-free and often entail substantial costs. For one, behavioral interventions entail some costs on the part of the government, to develop and implement them, although these costs tend to be more limited than the costs required to implement comparable traditional regulations that use subsidies, mandates, or bans.

Nudges also generate a myriad of private costs. Among them one can find direct cognitive judgment or decision costs for the targeted individuals, such as when a nudge requires or encourages people to pay greater attention to their choices, process more information, engage in more thorough deliberation, or even simply to make a choice they may wish to avoid. The same may also produce attendant emotional costs, as when the need to process further information or make a difficult decision causes internal emotional conflict and strain.

Quite naturally, however, the greatest emotional costs are often generated by nudges that intentionally recruit affect to impact people's behavior, as exemplified by the graphic warning labels mandated for cigarette packages in many countries (World Health Organization 2017).

Yet the most significant and often ignored category of costs concerns the private costs—most notably the opportunity costs that nudges generate when they change people's behavior, because of the inevitably forgone benefits of these individuals' former course of action. Like other regulation, nudges that change behavior entail opportunity costs even when they make people better off on balance. But a closer look reveals that behavioral instruments routinely make at least some of their targets worse off, thereby generating significant private opportunity costs (Tor 2020b, 2023). Finally, behavioral interventions occasionally generate costly spillovers, such as when individuals who engage in nudge-induced socially beneficial behavior (e.g., recycling) "self-license" to engage in other, socially costly, conduct (e.g., increase their lawn water use) (Thorgerson and Olander 2003).

Ideally, a better appreciation of the welfare effects of nudges—including their costs as well as their benefits—would enable scholars and regulators to conduct a full cost-benefit analysis (CBA) of behavioral interventions, just like in the case of traditional regulation (Boardman et al. 2018; Ellig et al. 2013). But when the inputs necessary for a full-fledged CBA are unavailable, other approaches may help determine the desirability of a given nudge or at least its relative appeal compared to competing regulatory instruments.

One familiar option is cost-effective analysis (CEA)—a common method of assessment that regulators also use when they are unable or unwilling to monetize policy costs (Layard and Glaister 1994). CEA generates a cost-effectiveness (CE) ratio by dividing policy costs by a non-monetary measure of their impact or effectiveness to enable comparisons of competing interventions in terms of their costs per each unit of effectiveness (Layard and Glaister 1994; Levin and McEwan 2001). Yet CEA is incapable of determining which competing policy is more efficient or even whether any available regulatory instrument is likely to provide net social benefits (Boardman et al. 2018). If that were not enough, CEA still requires the monetization of all policy costs, a challenge that policymakers are often loath to overcome.

Given the costs and challenges of conducting a full CBA, and CEA's many additional limitations, another method of nudge assessment—namely, rationality-effects analysis (REA)—has recently been proposed (Tor 2019, 2023). Instead of attempting to monetize the full range of a policy's costs—a hurdle that CBA and CEA must both overcome—REA focuses on the expected effects of the nudge on the rationality of its targets. This focus allows the analyst to distinguish, for instance,

rationality-promoting nudges that merit a presumption favoring their implementation from their rationality-diminishing counterparts that bear a presumption against their adoption.

Part 1 reviews the development of nudge definitions and explains what makes some definitions more useful than others, while Part 2 follows by highlighting some of the key mechanisms through which behavioral instruments change behavior and their significance for appropriate policy assessment. Part 3 describes the main reasons for the increasing employment of behavioral regulation around the globe and the developing evidence for nudge efficacy, after which Part 4 considers the myriad costs of behavioral regulation, emphasizing the most significant cost category among them—that is, the private opportunity costs of successful nudges. Part 5 draws on the preceding Parts and recent evidence to demonstrate cost-benefit analysis, cost-effectiveness analysis, and rationality-effects analysis of behavioral instruments.

2 Nudge Definitions[1]

2.1 Libertarian Paternalism and the Origins of Nudging

Behavioral regulation has received much attention over the last two decades, with a veritable flood of academic writing on the topic beginning in the late 2000s, on the heels of Richard Thaler and Cass Sunstein's 2008 book *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yet notwithstanding the frequency of its usage, "nudge" remains an ambiguous term, with different scholars taking it to mean different things, albeit usually relating to the employment of (some) behavioral or behaviorally-informed instruments to advance (some) policy goal.

Considering alternative nudge definitions and clarifying the term's usage is important not only because of the benefits of conceptual clarity. After all, this article describes behavioral regulation from a law and economics perspective and examines how we might assess the welfare effects of specific nudges. Having a more precise and coherent delineation of nudging therefore should help to distinguish it from other policy interventions, explain its appeal, and more clearly identify its various effects.

The earliest formulation of the nudge concept did not use that term. Instead, in their groundbreaking 2003 article, Sunstein and Thaler advocated for "libertarian paternalism". They argued that, since deviations from rationality lead people to

1 This Part draws on **Tor (2021a)**.

make suboptimal choices, appropriate interventions can make individuals better off by using an approach that "preserves freedom of choice but encourages both private and public institutions to steer people in directions that will promote their own welfare" (2003: 1201). In particular, Sunstein and Thaler (2003) proposed using behavioral instruments—like defaults, anchors, and framing effects—to encourage individual choices that improve private welfare. Hence, what renders an intervention libertarian paternalistic is the unique combination of its private welfare goal and its choice-preserving tools (Tor 2016).

It was Thaler and Sunstein's (2008) follow-up book that gave nudges their name and popularized libertarian paternalism. The book defines nudges as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any option or significantly changing their economic incentives" (2008: 11), so long as the intervention remains libertarian paternalistic (15). And since choice architecture in the authors' parlance simply means "the context in which people make decisions" (2008: 12), it would appear that nudging is just a slightly refined version of libertarian paternalism.

Despite the apparent near identical nature of nudging and libertarian paternalism in their creators' parlance, however, the 2008 adoption of the new, informal and intuitive, nudge terminology turned out to exert a profound effect on both Thaler and Sunstein's own policy recommendations and the broader discourse surrounding behavioral regulation that followed (cf. Rizzo and Whitman 2019). For one, because they feel they already know what a nudge is, few scholars pause to examine whether each behavioral intervention they discuss is a nudge in the specific sense advocated by Thaler and Sunstein (2008) or, indeed, in any well-delineated sense. Those who do attend to the meaning of nudging in their analyses often follow Thaler and Sunstein's (2008) approach (De Haan and Linde 2018; Oliver 2015).

In addition, the nudge usage focuses attention on the policy instrument alone, such that all "soft" behavioral instruments seem like nudges. Yet the softness of the policy tells us little about its choice-preserving (libertarian) credentials and even less about whether it is concerned with private welfare (paternalism) as opposed to public welfare. Therefore, the nudge usage obscures the distinction among libertarian paternalistic interventions and other behavioral policies that are paternalistic but non-libertarian, libertarian but non-paternalistic, or even neither libertarian nor paternalistic.

Given the tendency of nudge terminology usage to slip towards both non-libertarian or non-paternalistic interventions, one finds Thaler and Sunstein (2008) already applauding purported nudges that exceed both the libertarian and the paternalistic boundaries of libertarian paternalism. One such case concerns the authors' social nudge example of a regulation requiring firms to publish

"Toxic Release Inventories," which enable the media and environmental groups to more easily produce an "environmental blacklist" that threatens polluters with substantial social sanctions (2008: 190–191). Whatever the merits of these mandated inventories, however, their threatened sanctions obviously are not libertarian and the environmental goals they pursue are matters of public welfare, rather than paternalism.

Therefore, it should come as no surprise that other scholars find the term intuitively appealing but, possessing no particular commitment to its origins in libertarian paternalism, use nudging simply as a loose shorthand for policies with some behavioral component or connection, irrespective of their goals or the specific mechanisms through which they generate their effects (Sibony and Alemanno 2015). Examples of the three types of broader nudge use—namely, interventions that are non-libertarian but paternalistic, libertarian but non-paternalistic, or neither libertarian nor paternalistic—abound in the literature (see generally Tor 2021a).

Consequently, purported nudges sometimes lack a meaningful connection to the behavioral or behaviorally-informed methods, or to the justifications, of traditional nudges. At the extreme, some have even claimed traditional price instruments like taxes and subsidies as nudges, merely because they do not literally compel choice (Le Grand and New 2015). Most importantly for present purposes, however, overinclusive nudge usage obscures both the commonalities that many behavioral instruments share and the fundamental differences between them and traditional regulation in terms of their likely welfare effects.

2.2 Better Nudge Definitions

A number of scholars have noted the varied and inconsistent ways in which the nudge terminology has been employed following Thaler and Sunstein's (2008) introduction of the term (e.g., Sibony and Alemanno 2015). In particular, efforts at more precise nudge definitions have recently been made by economists who require tractable formulations of these instruments if they are to mathematically model or empirically test their welfare effects (Allcott and Kessler 2019; Bernheim and Taubinsky 2018; Farhi and Gabaix 2020; Spiegler 2015). These competing economic formulations overlap significantly, but still differ in important respects.

Allcott and Kessler (2019: 236), for one, define nudges as interventions that "affect choice without changing prices or choice sets" and explain that they intend their approach "to be consistent with the practical examples of Thaler and Sunstein (2008)" (2019: 241). Nonetheless, Allcott and Kessler (2019: 241) build their model to account for the possibility that "a nudge provides information, reduces bias, and/or persuades people by activating moral utility". These authors

therefore offer a more nuanced approach that accounts for at least some key behavioral channels through which nudges exert their effects, although their model still does not distinguish nudges from traditional non-price instruments.

Bernheim and Taubinsky (2018: 441) aim to rectify this overinclusiveness by defining nudges—which they call "non-standard" instruments—so as to distinguish them from standard non-price instruments like quantity regulation or information disclosure. Yet Bernheim and Taubinsky (2018: 442) further narrow their definition to include as a true nudge only "a non-price intervention that achieves a change in behavior by modifying the decision problem in a way that would not alter a consumer's perception of the opportunity set absent some error in reasoning". This narrower definition limits nudges to purely behavioral instruments that would not exert any effect on the choices of strictly rational consumers, thereby excluding many interventions routinely referred to as nudges that can affect opportunity sets. According to this formulation, for instance, social information policies (e.g., Allcott 2011; Allcott and Kessler 2019) or affect-laden interventions would not qualify as nudges, because they may contain some rationally-relevant information or exert some economic costs.

The narrow approach of Bernheim and Taubinsky (2018) clearly distinguishes nudges from traditional non-price instruments that change the opportunity set of strictly rational actors and attends to the mechanisms through which different policies change behavior. However, Bernheim and Taubinsky's (2018) approach implicitly treats behavioral policies that produce any objective price or opportunity set effects as if they were traditional instruments. In reality, however, these instruments often change behavior through multiple different mechanisms, some entailing objective opportunity set changes (e.g., emotional costs) while others do not (e.g., presenting objective information in a user-friendly format) (Allcott and Kessler 2019). Hence, while analytically defensible, this definition has limited practical usefulness for assessing the welfare effects of the many real-life policy instruments that possess both traditional and significant behavioral elements.

One solution to this limitation is to adopt a less restrictive definition that still defines nudging with attention to its behavioral mechanisms and their relationship to rationality. Specifically, Tor (2021b) defines nudges as *significantly behavioral instruments*—namely, policies whose impact is due, at least in significant part, to the activation of behavioral processes that rational actors would find irrelevant. This better definition avoids the problem of overinclusiveness, because it excludes all traditional instruments, including policies that merely aim to convey information. In turn, the "significantly" behavioral threshold of the definition responds to the underinclusiveness challenge by bringing into the nudge fold those prevalent real-world interventions whose effects are driven by some combination of behavioral and traditional factors, like the ubiquitous Home

Energy Reports (Allcott 2011; Allcott and Kessler 2019). So long as their behavioral effects are not merely minor byproducts of a traditional instrument, considering such policies as nudges helps direct analysts' attention to their behavioral as well as to their traditional elements.

3 Nudge Taxonomies: Organizing the Behavioral Toolbox[2]

Defining nudges as significantly behavioral instruments also highlights the centrality of the systematic differences between real, boundedly rational human behavior and hypothetical strict rationality for the proper understanding of nudging and its welfare effects. After all, bounded rationality is required to justify paternalistic policies, since the hypothetical rational actor always maximizes her welfare and thus cannot benefit from paternalistic interventions (Tor 2016; Zamir 1998). Our nudge definition also highlights the similarly critical role of bounded rationality in public welfare nudging that seeks to change private behavior in significant part by activating psychological processes that impact only boundedly rational individuals.

The classification of nudges is important not only because they shape the conduct of boundedly rationality individuals in many different ways but also because behavioral instruments can differ dramatically in their welfare effects, depending on the instruments they use, the psychological mechanisms they activate, the behaviors they target, and more.

3.1 Extant Nudge Taxonomies

Some nudge classifications emphasize the nature of the instruments—or "techniques" (Berthet and Ouvard 2019)—they employ, as exemplified by Johnson et al.'s (2012) many nudge categories, like reducing the number of alternatives, using technology and decision aids, setting defaults, adjusting the time frames and sequences of choices, partitioning options and attributes, and designing attributes. Munscher et al. (2015) classified nudging instruments into the three very broad categories of decision information, decision structure, and decision assistance (each of which encompassed further sub-categories with a resulting list of nine types of nudge techniques). And Hollands et al. (2017) offered a detailed, context-specific, technique-based taxonomy that more closely tied nudges to the concrete behavior changes they seek.

2 This section is based on **Tor (2019, 2021b)**.

The advantage of technique-based taxonomies lies in their focus on the observable aspects of nudging, but they may be limited in their descriptive power to explain or predict the effects of a specific nudge (Grune-Yanoff and Hertwig 2016; Tor 2023) and, when based primarily on conjecture regarding the likely impact of a given intervention, lacking a clear normative justification for the nudge's employment (Michie et al. 2008). In response to these limitations, some offered mixed taxonomies that connect the techniques of nudging with relevant psychological constructs these instruments aim to impact.

For example, Lin et al. (2017) distinguished between type 1 nudges—which are subtler, more likely to encourage intuitive or automatic response, and less likely to lead people to consider and reassess their behavior—and type 2 nudges, that "aim to promote a sustained reevaluation of the evidence base on which people make their choices, and the choices themselves" (296). In the context of nudging to battle unhealthy dietary habits, for instance, introducing smaller-sized plates would be a type 1 nudge, while menu calorie labeling is a type 2 nudge. Hence, Lin et al. (2017) categorize nudges based on the psychological processes they seek to activate and designate all interventions that recruit automatic processes as type 1 nudges.

Lin et al. (2017) take a step forward in the categorization of behavioral instruments by directly linking nudge techniques with general types of psychological processes. No less importantly, they demonstrate how the usefulness of a given taxonomy depends on its purpose. Thus, Lin et al.'s (2017) categorization provided a basis for their assessment of the relative efficacy of nudges that demand greater attention and deliberation compared to policies that minimize such demands and largely circumvent conscious information processing.

Nevertheless, the type 1/type 2 distinction may not suffice for evaluating other aspects of nudging beyond its efficacy. For this reason, some scholars put forth frameworks that reflect their specific concerns, like the philosophical criticism that nudges can infringe on individual autonomy (Barton and Grune-Yanoff 2015). Authors who share this concern naturally classify nudge techniques based on their autonomy effects, as illustrated by Baldwin's (2014) "three degrees of nudging" taxonomy. Others suggested classifications based on normative considerations like freedom, political legitimacy, and more. Barton and Grune-Yanoff (2015), for instance, distinguish among interventions that employ automatic processes (heuristics), those that block some undesirable operations of such heuristics, and policies that simply provide useful information without directly impacting heuristics. Here, again, the proposed taxonomy is driven by specific normative considerations: Barton and Grune-Yanoff (2015) are more concerned about nudges that trigger heuristics than with interventions that block them or provide information.

In sum, most recent taxonomies recognize some type 1/type 2 distinction, while adapting this general concept to their purposes (Berthet and Ouvard 2019). In particular, researchers concerned with the normative evaluation of nudges may distinguish among different categories of the automatic processes they trigger, as illustrated by Baldwin's (2014) separation of type 1-like nudges that mainly rely on inertia or inattention (but can be resisted through reflection, according to the author) from policies that people would find difficult to overcome even upon reflection.

3.2 A Rationality-Based Taxonomy

Whatever their other merits, extant taxonomies do not distinguish nudges from one another based on their likely welfare effects—the normative consideration that matters most from an economic perspective. To address this shortcoming, Tor (2019, 2021b) developed a welfare-relevant taxonomy that differentiates among nudge techniques based on their rationality effects. The explicit linking of rationality and welfare extends the logic of economists' nudge definitions discussed in Part 1.2 to the classification of nudges. It recognizes that society tends to benefit more from instruments that promote individual rationality than from interventions that diminish it.

The role of rationality is most apparent in the case of paternalistic policies, which are needed only insofar as people fail to maximize their private welfare. Nudges that increase rationality make it more likely that people will act in their personal best interests, while those that diminish rationality must rely on policymakers' judgments of the matters, with their substantial attendant risks and costs (Tor 2016, 2019). Paternalistic nudges that increase rationality also tend to generate positive spillover effects while rationality-diminishing interventions tend to produce negative spillovers. To illustrate, people who are nudged to overcome some bias in their risk assessment (e.g., of contracting a disease) in one context (like vaccination) may exhibit a positive spillover by using their more rational assessment of that risk to make better decisions in other contexts (e.g., employing other protective measures or engaging in beneficial activities that entail increased exposure). The reverse holds for rationality-diminishing nudges: Even when they successfully lead individuals to take beneficial action in one context (such as causing them to vaccinate by making them overestimate their risk of contracting a disease), they may generate negative spillovers (like avoiding important medical procedures).

Rationality-increasing nudges with public welfare goals also tend make society better off than their rationality-diminishing counterparts. For one, private welfare benefits often constitute the majority of the overall benefits of public

welfare interventions generally (Allcott and Taubinsky 2015) and nudges in particular (Allcott and Kessler 2019; Tor and Klick 2022). In addition, the prevalence of spillover effects further militates in favor of rationality-increasing instruments and against rationality-diminishing ones, and more rational individuals tend to increase public welfare more than their less rational peers when exercising their political power at a local or national level (cf. Klick and Mitchell 2006).

In theory, increased rationality might risk producing negative externalities as well (e.g., when a more accurate assessment of the likelihood of enforcement suggests that legal violations are privately beneficial). However, not only are regulators unlikely to implement nudges that tend to produce such harms, but they also have at their disposal familiar, traditional instruments (like mandates or fines) that reduce the risk of negative externalities from more rational behavior on those occasions in which they turn out to be substantial.

Of course, significantly behavioral instruments are not limited only to rationality-promoting and rationality-diminishing nudges. Most are somewhere between these two extremes, ranging from policies that merely enable individuals to act more rationally (rather than become more rational), through rationality-neutral interventions that may change behavior without increasing or decreasing the rationality of their targets or their actions, to common techniques that exploit bounded rationality without rendering their target less rational.

Hence, the rationality-based taxonomy of nudges extends the logic of the familiar type 1/type 2 dichotomy. Its focus on rationality effects recognizes that although type 1 nudges—which primarily activate more intuitive, automatic processes—have a propensity to harm individual and social welfare, they are not all are equally harmful. Similarly, while type-2 instruments—which primarily activate more conscious, deliberative processes—are more likely to produce welfare benefits or, at least, avoid substantial harm to welfare, they are neither equally nor always beneficial. This taxonomy thus accommodates the broad range of welfare-relevant rationality effects of nudging, which depend not only on the specific instruments employed but also on their concrete contours and implementation context.

Rationality-promoting and rationality-enabling instruments typically call on type-2 deliberative processes and are therefore less likely to be harmful to individual or social welfare, and the same holds for rationality-neutral nudges, if to a lesser extent. Rationality-promoting techniques debias decision makers, improving their ability to advance their private welfare. For instance, a rationality-promoting policy might teach people more accurately to assess the relative risks of different activities or to avoid myopic decisions that contradict their own long-term preferences. The overall costs of rationality-promoting nudges may still exceed their benefits, in which case they are undesirable, but as a class they are the

most benign interventions available. Yet such policies are uncommon in practice (Tor 2019), since debiasing is difficult to implement at scale in regulatory contexts (Tor 2008).

Rather than increasing people's capacity for rational action, rationality-enabling nudges improve the environment within which people make their judgments or decisions. These instruments are popular with regulators because improvements to the decision environment—e.g., through simplification or the use of plain language—are relatively easy to implement and are viewed as uncontroversial. While potentially beneficial, however, rationality-enabling techniques also carry with them common risks, such as when they direct attention to one particular aspect of the available information that regulators believe will improve judgments or decisions (e.g., roadside signs highlighting year-to-date counts of roadside fatalities to encourage more careful driving). Consequently, although they encourage deliberative processes, these nudges risk distorting people's choices by diverting their attention from other important information (such as actual road risks; Hall and Madsen 2021). More generally, rationality-enabling interventions build on regulators' judgments of which information is most important and are thus capable of steering people towards welfare-diminishing actions.

The same holds to an even greater degree in the case of rationality-neutral nudges, which are benign in some cases but may generate welfare losses. These policies neither impact the rationality of people's behavior nor do they exploit their bounded rationality. Instead, they simply encourage individuals to act, consciously and deliberately, in ways that policymakers deem beneficial. Typical examples include active choice policies, certain reminders, and more. Notably, though rationality-neutral policies do not set out to change or distort individuals' beliefs or preferences but rather to engage type-2 processes, they nevertheless risk producing such effects on occasion. To illustrate, an active choice intervention may help some people beneficially avoid procrastination but pressure others to decide before they are ready to do so (Sunstein 2014). Similarly, a reminder that encourages people to take an action that will make them better off in the regulator's opinion (e.g., pay credit card debt on time to avoid late fees) may in fact produce net costs for them by erroneously implying that the encouraged action is their best option (i.e., better than paying late fees but avoiding higher overdraft fees; cf. Medina 2021).

Just as not all type-2 nudges merit the same deference, not all type-1 policies are equally or always harmful. Many popular behavioral instruments—from the setting of defaults, to framing, using order effects, and more exploit the bounded rationality of their targets without actively diminishing their

rationality. These interventions do not prevent people from engaging in deliberation, but nevertheless facilitate more automatic processing that renders careful analysis less likely or that is biased towards regulators' desired outcomes (cf. Baldwin 2014). Bounded-rationality exploiting nudges can entail substantial costs and spillover effects, yet may still produce social benefits, such as when they complement highly-beneficial hard instruments (e.g., facilitating compliance with criminal law). Therefore, while deserving of careful scrutiny, bounded-rationality exploiting policies occasionally may be desirable on balance.

In contrast, techniques that actively diminish rationality are almost universally undesirable. These nudges not only activate primarily intuitive, automatic processes, but aim to manipulate their targets' judgments and decisions. When regulators employ instruments like priming or anchoring, people may not even be aware that they are being nudged. But even if they are transparent—as when policymakers implement affect-laden policies (e.g., graphic warning labels)—rationality-diminishing nudging tends to be harmful. It produces substantial private costs and negative spillovers and its harms are difficult to avoid.

In sum, as our later discussion of nudge costs in part 4 and the assessment of nudges in part 5 will further demonstrate, a rationality-based taxonomy can help account for some of the main welfare effects of behavioral instruments.

4 The Benefits of Nudging

The increasing use of behavioral interventions by governments and other organizations reveals the attractiveness of these instruments for policy makers, whose enthusiasm can be attributed to the confluence of a number of factors. Section 3.1 discusses these factors, followed by Section 3.2's review of the emerging evidence on nudges' effectiveness—that is, the degree to which behavioral instruments in fact succeed in changing behavior in the field.

4.1 Why Regulators Like Nudges

Regulators like nudges for a number of related reasons: First, nudges are based on a realistic view of human behavior that is intuitively appealing; second and related, policy makers may believe that nudges are politically more feasible than their alternatives; third, the great variety of behavioral tools means that they are more versatile than traditional instruments; fourth, and finally, nudges tend to entail relatively low implementation costs, imposing less strain on limited budgets and thereby appearing to be more efficient or cost-effective than competing traditional instruments.

Policy makers find the realistic view of human behavior intuitively appealing, particularly when compared to the assumptions of rationality that undergird traditional economic models of regulation. David Halpern, a behavioral scientist who headed the United Kingdom's Behavioural Insights Team—the first official government unit dedicated to nudging—describes this appeal in his book *Inside the Nudge Unit* (2015: 7), writing that: "classic economic and regulatory models are themselves based on naïve models of humanity that do not ring true. They're like ill-fitting suits, because the model on which they are based is a simplistic mental mannequin. What would the world look like, and the actions of governments, businesses and communities, if based on *a more realistic model of people*?".

Being a sophisticated commentator, Halpern (2015: 7) concedes that a "practical approach to government based on a realistic model of people would be messier than traditional economics or law. It would need to reflect the complexity of the human mind—what we do well, and what we don't. When we design services and products, we would need to be respectful of this reality. We would have to design everything we do around people, not expect people to have to redesign their lives around us". Nevertheless, in the spirit of Thaler and Sunstein (2008), he argues that "[w]e are perpetually bombarded by subtle influences and cues, and nearly all of us—whether we like it or not—are at least "accidental nudgers" of a sort. The way a shop is laid out; how an offer is presented; how a form is written—all will have some kind of influence on behavior. In this sense, the world of nudging is all around us. *The question is, do we stumble on blindly, or seek to understand these influences and choices?*" (Halpern 2015: 7, emphasis added).

Halpern's (2015) rhetoric and the important question of whether governments can or should avoid nudging aside, comparable sentiments regarding the appeal of policy making that is based on a more realistic view of human behavior are voiced by policy makers in other countries as well. Perhaps most strikingly, former U.S. President Obama's Executive Order 13,707 (2015)—titled "using behavioral science insights to better serve the American people"—explicitly links the empirical foundations of behavioral research to the attractiveness of nudging, stating that: "[a] growing body of evidence demonstrates that behavioral science insights—research findings from fields such as behavioral economics and psychology about how people make decisions and act on them—can be used to design government policies to better serve the American people". Similarly, one of the earlier OECD reports surveying the state of behavioral policy making through 2016 (2018: 16) noted that "[t]his use of behavioural sciences has become commonplace in many countries to help institutions better design, implement, and enhance market interventions through factoring behaviour across a range of topics such as consumer protection, energy, environment, health, finance, and

taxation, among others". It is thus apparent that policy makers find nudges attractive at least in part because they believe that government interventions are better able to achieve their goals when based on a more realistic account of human behavior.

While their more realistic foundations make nudges attractive, public choice research shows that the actions of political actors and regulators alike are affected by their beliefs regarding the preferences of their constituents (e.g., Mueller 2003; Peltzman 1976). As with respect to other regulatory interventions, policy makers are less inclined to introduce nudges they expect to face substantial public resistance and more likely to employ them when they predict public acceptance. Sunstein et al. (2019: 1419) note, for example, that "[i]n democratic nations it is important to know whether members of the public will endorse such instruments."

One source of suggestive evidence of regulators' beliefs regarding the acceptability of nudges to the citizens they target is the behavioral turn in public policy around the globe. This turn is manifested not only in the large number of interventions across most policy areas that draw on behavioral insights and use behavioral instruments, but also in the swelling ranks of governmental "nudge units"—regulatory offices that are dedicated to developing, testing, and implementing behavioral policies together with other relevant government offices (e.g., Sunstein 2019).

The growing empirical literature on the extent to which individuals in different countries find behavioral regulation acceptable and the various personal, social, cultural, and political factors that affect nudge acceptability offers further evidence. After all, as Reisch and Sunstein (2016: 311) note, "survey responses provide relevant information, not least because public officials are inevitably responsive to what people think. If an intervention would trigger widespread public alarm, officials would be less likely to support it. In contrast, public approval can serve as a kind of permission slip". These surveys and experimental tests indicate that large segments of the public—often a majority—in many democratic nations find many nudges acceptable (e.g., Hagmann et al. 2019; Jung and Mellers 2016; Reisch and Sunstein 2016; Sunstein 2016; Sunstein et al. 2019), although support levels vary substantially across countries and types of nudges, and depend on a variety of factors (e.g., Arad and Rubinstein 2018; Bang et al. 2018; Gold et al. 2020). Thus, while this literature reveals factors that render behavioral policies more or less palatable to large swaths of their target populations, the emerging picture of significant public acceptance can help explain policy makers' enthusiastic endorsement of behavioral instruments.

Beyond their reality-based nature and public acceptance, regulators may also like nudges because of their versatility. Traditional policy instruments shape people's behavior primarily by restricting the range of options available to them

through mandates or bans or by changing the prices of these options by taxing or subsidizing them.[3] Unlike such blunt mechanisms of behavior change, the plethora of behavioral tools can modify many additional aspects of their targets' decision environment beyond the number of options available to them or their respective prices. To illustrate, nudges can display the same substantive information in different ways—using different emphases, frames, or ordering of alternatives; changing the structure of the decision setting by setting default options and arrangements, organizing the physical space or the design of objects that are relevant to the decision; and encouraging decision making and choice preservation through tools like required active choosing, reminders, or commitment devices (cf. Munscher et al. 2015). In all of these cases, behavioral instruments can shape people's behavior without making substantive changes to the choice set they face, thereby demonstrating a major appeal of nudging (Thaler and Sunstein 2008).

Last but not least, behavioral interventions attract policy makers in large part due to their promise of achieving policy goals at low implementation costs, allowing them to do more with limited government budgets. As Sibony and Alemanno (2015: 2–3) explain, "nudging is presented as a cheap and smart alternative to expensive traditional regulatory measures." Similarly, Sunstein and Reisch (2019: 3) state that "[t]he reason for the mounting interest [in nudging] should not be obscure. Nations would like to make progress on pressing social problems with tools that actually work and that do not cost a great deal." Other scholarship elaborates this point further, explaining that ""nudges" are cheap and easy to implement, because they allow to avoid (i) the direct cost of changing people's economic incentives and/or limiting people's action space, (ii) the monitoring cost of finding out which choice each individual made and, possibly, the cost of punishing or rewarding each choice, and (iii) the technical difficulties associated with finding out individual choices" (Capraro et al. 2019: 1). The low implementation costs of behavioral policies thus offer policy makers an important budgetary advantage from a bureaucratic perspective.

Nudges' low implementation costs also appear to make them more cost-effective than traditional instruments, as influentially argued by a group of prominent scholars, who reviewed extant studies in some policy areas in which evidence on nudge efficacy was available, including retirement savings, energy consumption, adult outpatient influenza vaccinations, and more (Benartzi et al. 2017).

3 Two important exceptions to these main effects of traditional instruments, which merit separate analyses, concern traditional policies that mandate the disclosure of private information (rather than merely providing access to costly public information) and the law's ability to change people's beliefs and transform their preferences.

In each area, these researchers assessed the effectiveness and implementation costs of different interventions, producing an effectiveness-cost (EC) ratio to measure the relative effectiveness of behavioral versus traditional (primarily financial) interventions. Benartzi et al. (2017) found that the most effective nudges held a consistent, substantial relative effectiveness advantage, which was nearly always attributable to their substantially lower implementation costs (Tor 2023; Tor and Klick 2022).

All in all, a number of factors make nudges attractive to policy makers, from their more realistic behavioral foundations, through their substantial public acceptability and great versatility, to the low implementation costs they entail for the government. Notwithstanding their appeal to regulators, however, to be effective behavioral policies must also produce private or social welfare benefits.

4.2 Nudge Efficacy

The empirical evidence documenting the effectiveness of behavioral policy interventions in the field is limited but rapidly growing. Early nudging studies developed out of the extensive empirical research that documents how the design of decision problems and information or the decision context impact behavioral outcomes. This research was conducted primarily in controlled laboratory settings, but additional studies have often been performed in the field, with relevant populations, the better to assess the effectiveness of nudges as real-world policy instruments.

Recent reviews based on academic publications show that nudges have already received some testing, mainly with respect to private welfare interventions in domains including consumer choice, education, finance, and health, but public welfare policies in the areas of environmental protection and sustainability, prosocial behavior, and more (Hummel and Maedche 2019; Szaszi et al. 2018) have also been examined. In the academic literature, nudges have been studied most extensively in health research, often focusing on dietary behavior (Bauer and Reisch 2019; Vecchio and Cavallo 2019), but also evaluating other health-related activities, like self-management by patients with chronic diseases (Mollenkamp et al. 2019) or the promotion of physical activity in the general population (Forberger et al. 2019).

Broad overviews find some nudges capable of producing behavior change while others prove far less effective. With respect to private welfare nudges, for example, a summary of thirty-nine literature reviews and meta-analyses of behavioral interventions to improve dietary choices—a comprehensive assessment of all relevant publications between 2010 and 2017—reported that "virtually all reviews" found that "nudges hold promise in fostering healthier food choices"

(Bauer and Reisch 2019: 14). At the same time, the substantial differences among the tested interventions in terms of the specific instruments they employed, their settings, and the quality of their designs, repeatedly precluded researchers from drawing general conclusions about nudge effectiveness in the health domain (Bauer and Reisch 2019; Vecchio and Cavallo 2019).

This general pattern is concretely illustrated by the meta-analysis of Arno and Thomas (2016), who assessed nudge effectiveness in improving adult dietary behavior based on thirty-seven publications that encompassed forty-two independent studies. The authors found that, on average, these interventions caused a 15.3% increase in healthier consumption decisions as measured by the frequency of healthy choices or by overall healthy food intake. Notably, however, although this result demonstrates a substantial average increase in the desired behavior, the effects of different nudges were highly variable, with about a third of the reported studies showing only a small positive effect and occasionally even a negative effect, even while nearly a quarter of the studies reported very large effects (of 30–50%, even after the exclusion of two outliers that reported increases of 79% and 129% respectively).

A similar picture emerges with respect to policies encouraging pro-environmental behavior—the most common public welfare nudging area. For instance, Byerly et al. (2018) reviewed 72 studies that tested 160 different interventions—comparing the effects of nudges to those of educational and incentive-based interventions—using a broad definition of pro-environmental policies that covered areas ranging from family planning and meat consumption, through transportation choices and land management, to waste production and water use. While noting that only a small portion of their sample included direct comparisons of competing instruments, Byerly et al. (2018: 166) concluded that "[o]verall, contextual interventions outperform education interventions. However, it is less clear how contextual interventions compared to financial incentives." In addition, while finding that some nudges produced significant effects, the authors cautioned that the effectiveness of behavioral instruments often depends on factors such as the personal characteristics of their targets, the context of the intervention, and more, thereby indicated they are unlikely to be universally effective.

Following these and similar findings regarding the heterogeneity of nudge effects, a quantitative review by Hummel and Maedche (2019) sought *inter alia* to compare the effectiveness of different behavioral instruments and to assess the relative importance of the particular context of both the intervention and the specific type of nudge it employs. The authors were able to identify 100 higher-quality primary publications with 317 independent effect sizes spanning policy areas including health, finances, the environment, energy use, and more, that

reported sufficient statistical information for quantitative comparisons. Hummel and Maedche (2019) found about one-third (118) of the policies failed to reach statistical significance, while the remainder (190 interventions) were nearly evenly split between low, medium, and high relative effect sizes—respectively defined as less than 10%, 10–30%, and more than 30%. Overall, the nudges in the sample had a median relative effect size of 21% and an average effect size of 30% (after excluding outliers), with the effects in specific studies ranging widely, from 0% to 1681%.[4]

When comparing effectiveness in major policy areas, Hummel and Maedche (2019) reported that behavioral interventions were most effective in the domains of privacy (with a median effect size of 44%) and the environment (39%), least effective in the energy use category (13%), and intermediate for the finance (28%) and health (21%) areas. The variability in the effect size of different nudge types was more dramatic, however: Defaults, the most common and most effective behavioral instrument in the reviewed literature, showed a large median effect size of 50%, while that of simplification—the next most common nudge category—was only 20%. Moreover, reminders and precommitments, for instance, produced only small median effect sizes of 8% and 7% respectively. Overall, therefore, this broad quantitative assessment of published research clearly shows that certain nudges are far more effective than others in changing their targets' behavior.

Finally, an important recent contribution by DellaVigna and Linos (2020) provides further insight into the effectiveness of real-world nudging by comparing the results of meta-analyses of behavioral interventions in research studies (like those assessed in the reviews of the academic literature discussed above) with those documented for large-scale policies implemented by two governmental "nudge units" in the United States—Behavioral Insights Team North America (that operates with local governments) and the Federal Office of Evaluation Sciences. The latter data are unique, based on comprehensive records of all interventions conducted by the two units for about 4 years (2015-July 2019) totaling 165 trials with 349 different nudge treatments targeted at over 37 million participants. Importantly, the nudge-unit data is not only based on large-scale field interventions, but also includes many treatments that have not been published in academic outlets and thus have not been subjected to the usual selection effects that strongly favor the publication of research with statistically significant results. Hence, both

4 One should bear in mind, however, that small absolute changes in a dependent variable can generate very large *relative* effect sizes, as when the warning nudge of Khern-am-nuai et al. (2017) increased participants' password strength scores from 0.0054 to 0.0962, thereby producing a relative change of 1681%.

its scale and its unbiased nature make this nudge-unit dataset highly informative regarding the effectiveness of behavioral policies in the field.

After narrowing down the dataset to render the included interventions more comparable to one another, DellaVigna and Linos (2020) retained a final sample of 126 randomized controlled trials (RCTs) involving 243 nudges and over 23 million target participants, which they compared to a similar subsample from the set of academic studies that Hummel and Maedche (2019) reported on, which consisted of 26 RCTs with 74 nudges and more than 500,000 participants. Notably, DellaVigna and Linos (2020) found that the nudges in their academic subsample produced an average relative effect size increase of 33.5% in the desired behavior over the 26.0% baseline of the control groups, or an 8.7% average absolute increase in the frequency of the nudged behavior. On the other hand, the nudge unit data showed a dramatically smaller average relative effect size increase of 8.1% from a 17.2% control baseline, or a 1.4% average absolute increase in the frequency of the nudged behavior—that is, only about one-sixth of the magnitude of the academic subsample effect size.

DellaVigna and Linos (2020) attribute the striking difference between the two subsamples to two main sets of factors: First, the subsamples differed substantially in their statistical power and exposure to selection effects. The nudge unit interventions targeted far larger participant groups than those available to the academic studies, providing the former with far greater statistical power that enabled them to identify significant effects at smaller effect sizes than the latter required; related, because academic journals usually publish only statistically significant findings, the design, submission, and acceptance processes of the academic studies were likely to generate a selection bias that combined with the studies' lower statistical power to favor nudges that produce larger effect sizes.

Second, the characteristics of the nudges in the two samples differed systematically in terms of their medium of implementation (more in person for the academic studies versus non-interactive means, such as email or letters for the nudge units); the policy areas on which they focused (e.g., with academic nudges emphasizing the domain of health with considerable attention to the environment, while the latter domain was virtually absent from the nudge units sample, which focused most on revenue and debt); and the techniques they used (with the academic studies drawing far less on simplification and personal motivation and more on choice design than the nudge unit policies).

In sum, DellaVigna and Linos's (2020) findings are highly informative, suggesting that nudges clearly can be effective when implemented at scale, though the magnitude of their effects under these circumstances may often be small in absolute terms even if statistically significant. Further analyses by these authors also concluded that the 1.4% absolute effect size they obtained for the nudge unit

interventions is a robust and reliable estimate of their performance. While these findings provide a valuable benchmark, however, further extrapolation from them should be done with care, particularly when considering those common nudges whose characteristics resemble the academic subsample more closely than the nudge unit subsample (e.g., employing more personal or interactive approaches, advancing environmental policies, or using choice design). Moreover, DellaVigna and Linos (2020) explicitly excluded from their analysis default nudges—which Hummel and Maedche (2019) found most common and most effective—raising the likely possibility that default-based nudges would outperform the study's benchmark.

Finally, the nudge units' interventions were necessarily subject to time and political constraints, which led to the implementation of less controversial policies that were easier or quicker to implement (cf. Halpern 2015), as manifested by the medium, policy area, and instrument choices they made. Besides implying, as already noted, that future interventions that deviate from these characteristics may prove more effective, these nudge unit policies are notable in rarely constituting standalone behavioral instruments. Instead, in the main policy area of revenue and debt, for instance, a typical simplification nudge would seek to facilitate individual compliance with extant rules and regulations (e.g., a tax). The same pattern holds for the second most common nudge unit policy area of benefits and programs, in which behavioral interventions try to encourage the uptake of extant government benefits. While prevalent and practically important in some settings, however, this type of complementary, "add on", nudge may perform very differently from standalone instruments—like most of those tested in the academic sample studies (e.g., Home Energy Reports)—that seek to change people's behavior in the absence of a mandate or a complementary financial incentive policy.

5 Nudge Costs

Like other regulation, behavioral policies entail both public and private costs. As already noted, however, the implementation costs of nudges tend to be smaller than those of traditional instruments. On the private side of the ledger, some nudge costs are borne by all or many of their targets and even by third parties. Often, however, the most significant nudge costs are the private opportunity costs these policies entail for the individuals whose behavior they successfully change. Sections 4.1 and 4.2 therefore briefly discuss the implementation costs of behavioral instruments and their private costs respectively, while Section 4.3 examines their private opportunity costs in greater depth.

5.1 Government Implementation Costs

Nudges typically entail lower implementation costs than comparable traditional instruments, since the latter usually seek universal compliance by the targets and require enforcement. For instance, a law mandating that drivers wear a seatbelt requires the investment of significant regulatory resources in policing drivers to deter violations, identifying violators and prosecuting them, and adjudicating disputed violations (apart from any costs incurred by drivers). These costs are avoided, however, if law makers instead encourage seatbelt use via nudging, such as by requiring car manufacturers to install an automatic alarm that is triggered when a driver starts a vehicle without fastening her seatbelt, but which the driver can also turn off at will. Manufacturers still need to design and install this alarm, but the per-driver implementation costs of this policy are small, infrequently incurred, and a mere fraction of the implementation costs of the ongoing enforcement of a seatbelt mandate.

Despite being non-coercive, the implementation of traditional financial interventions, such as taxes or subsidies, can also be quite costly for the government. Financial instruments seek to facilitate a change in their targets' behavior using positive or negative financial incentives that entail a significant budgetary price tag. For example, when law makers can offer tax deductions to facilitate charitable donations, the more successful these deductions are in increasing donations the more they diminish the government's tax revenue. In contrast, a policy that successfully increases donation rates through purely behavioral means—such as by encouraging or emphasizing social norms that favor donations (e.g., Zarghamee et al. 2017)—entails dramatically lower implementation costs. No less importantly, the tax system entails large adminstrative costs for the government to administer, audit, enforce, and so on, most of which are absent when purely behavioral instruments are employed instead.

The implementation cost advantage of nudging is clearly illustrated by Benartzi et al.'s (2017) estimates of these costs, comparing behavioral to traditional instruments across a number of key policy areas. For example, in the domain of retirement saving contributions, the authors found that a requested active choosing nudge studied by Caroll et al. (2009) cost merely $2 per employee to implement, by preparing a form to distribute to the employees and a follow up phone call to those among them who failed to make the requested choice. In contrast, Benartzi et al. (2017) assessed the implementation costs of competing traditional interventions as ranging from a low of $4.04 per employee for a program tested by Duflo and Saez (2003), which provided some employees incentives to participate an educational session explaining the benefits of a retirement savings program, to a high of $195 per person affected for a Danish tax law change

that modified the tax benefits associated with a retirement savings vehicle, studied by Chetty et al. (2014). In fact, the lowest traditional policy implementation cost estimate of $4.04 probably understates the case somewhat, since it ignores the per-employee cost of the educational program even while including the effects of the intervention on employees who did not receive the financial incentive but were in the same department as their incentivized peers.

A similar pattern holds for each of the three additional policy domains that Benartzi et al. (2017) examined. The authors estimated the implementation costs of a nudge to encourage college enrollment among recent high school graduates at $53.02 per program participant for training of and payment for tax professionals, materials, software, and call-center support, and those of two competing financial incentive programs at $4,468 per college student for subsidies in one case and $5,181 per eligible person for stipends in another. They also estimated the implementation costs of two energy conservation nudges at $1 per report (with reports sent monthly, bimonthly, or quarterly) and $3.02 per household, respectively, and those of two competing traditional programs at $5.09 per customer ($3.70 for rebates plus $1.39 for administrative and marketing costs) and $10.83 per customer on average (for financial incentives and education). Finally, Benartzi et al. (2017) estimated the costs of two nudges targeting influenza vaccination at $0.33 per employee (for adding planning prompts to reminder letters) and $3.21 per person (for unutilized clinic capacity) respectively, and those of the two competing traditional interventions at $6.03 per eligible student (for a monetary incentive) for one and $15.21 per employee for the other (including $0.93 of educational cost plus a free vaccine cost of $14.28).

Finally, while more recent work shows that Benartzi et al.'s (2017) figures overstate the implementation costs of traditional financial instruments, the observation that nudges usually entail modest costs for the government. Tor and Klick (2022) explain that the economic transfers involved in financial instruments do not constitute true economic costs. As these authors demonstrate,once such transfers are excluded from the implementation cost figures, financial incentive policies turn out to be far less costly then they appear when transfers are erroneously included (Tor and Klick 2022). Although this important correction diminishes the seemingly dramatic advantage of nudges over traditional financial instruments, however, it does not change the fact that behavioral instruments typically entail only modest implementation costs.

5.2 Private Costs

Most nudges, but particularly rationality-enabling or rationality-neutral nudges that primarily activate system 2 processes, can impose on their targets' judgment

or decision costs, by directing people to pay greater attention to their choices, process more information, engage in a more thorough deliberation, or even simply to make a choice they would have avoided but for the nudge.[5] Consider, for example, Carroll et al.'s (2009) required active choice study, in which employees were asked to choose their preferred retirement savings contribution rate but left them free to decide whether to join the plan and how much to contribute (imposing no penalty on those who failed to make the choice). Aside from its other benefits and costs, this nudge imposed on all new hires the cognitive and time costs required to read the form and grapple immediately upon hiring with the significant decision of whether and how much to contribute to their retirement savings. These costs may have been meaningful not only for the 28% of employees who were successfully nudged to join the plan but also for the 31% among them who actively chose not to join it and thus obtained no benefit from the intervention. For these participants, the decision may well have entailed cognitive costs to process all the relevant information, other psychological and emotional costs associated with making a difficult tradeoff between savings and consumption, and the economic costs of the time spent over the decision (Janis and Mann 1977; Sunstein 2014; Weber et al. 2001).

Another active choice variant further illustrates how a nudge—in this case one that combines rationality-neutral and bounded rationality exploiting elements—can impose on their targets not only decision costs but also incidental, emotional costs. In a series of controlled experiments, Keller et al. (2011) tested "enhanced active choice" interventions that not only asked their targets to choose but also formulated the available options to highlight the costs of not making the choice favored by policy makers. This "enhancement", which the studies found effective, was meant to use participants' loss aversion to nudge them further towards a particular choice, and the authors also found some evidence that their effect was partly mediated by their targets' increased regret aversion when asked to make an explicit choice.

In fact, many common nudges impose direct emotional costs incidentally or as a means for encouraging behavior change, an "emotional tax" that can reduce consumer welfare without generating government revenues (e.g., Glaeser 2006). Two studies of donation behavior by Damgaard and Gravert (2018) demonstrate how such costs can be imposed by relatively benign instruments, such as mere reminders sent to potential donors who previously provided their email address

5 The potential significance of decision costs is also illustrated by the nudges that succeed by lowering their targets' decision costs, such as BIT's (2012) finding in a U.K. energy savings intervention that reducing households' decision costs for insulating lofts by merely introducing a combined offer of loft insulation and cleaning can increase take-up.

to a charity. In the first study, the reminder increased the number of actual donors but also increased the rate at which potential donors unsubscribed from the email list. The second study further estimated an average reminder "annoyance cost" at approximately $2. Moreover, because only 1.2% of those on the list donated in any given month, this small cost nearly offset the nearly hundred-fold greater estimated "warm glow" benefit to actual donors, so that the average net private benefit to list members was only about $0.22.[6]

Social information interventions, which provide their targets with social comparisons as well as information regarding actual or purported social norms, also impose emotional costs on some of their targets (Tor 2023). For instance, Allcott and Kessler's (2019) extensive field study of the welfare effects of Home Energy Reports (HERs)—a ubiquitous social information nudge seeking to encourage energy conservation—found a substantial majority of recipients (59%) exhibiting a negative willingness to pay for the reports. These participants valued the reports negatively enough that the average direct costs to program participants were comparable in magnitude to the program's implementation costs.

The emotional costs of behavioral instruments are likely to be even more notable for policies that intentionally recruit affect to impact behavior. One familiar intervention on point is the extensive use of graphic warning labels on cigarette packaging. The World Health Organization (WHO 2017) considers these labels the most effective tool for tobacco control. Unsurprisingly, insofar as the labels rely on their targets' emotional reactions to try to change their behavior, a meta-analysis of 37 experimental studies of pictorial cigarette pack warnings found they produced stronger negative emotional reactions, such as fear or disgust, than text warnings (Noar et al. 2016).

In addition, nudges can also produce social costs, particularly for those who resist them. At the most basic levels, individuals who refuse to follow a popular nudge may receive social disapprobation or even social sanctions for failing to conform, much like those who violate social norms (e.g., Morris et al. 2015; Legros and Cislaghi 2020). Such social costs are more likely for nudges that publicly highlight individuals' performance on a socially-relevant metric, as illustrated dramatically by Butera et al.'s (2021) "public recognition" interventions. In these authors' first field study at a YMCA the nudge revealed each individual participant's attendance and donation amount to all other participants. Their second, online, experiment used an even stronger manipulation, in which participants' contributions to the Red Cross were publicly shared with others in the experiment through a webpage

6 Thunström (2019) similarly finds that informational nudges that merely make some information more salient to consumers can impose emotional costs (as well as benefits) on their targets, particularly on those who do not respond to the intervention.

that posted individuals' photos, the amount they raised, their rank relative to other participants, and (for two of three subject pools) the participants' names. Unsurprisingly, Butera et al. (2021) found less than 27% of participants indifferent to their public recognition manipulation, with an even smaller proportion—of merely 7% and 11% respectively—in the two samples in which participants were likely to know or recognize one another.

Besides their cognitive and emotional costs, nudges may also generate some direct economic costs. For example, behavioral instruments that facilitate deliberation also require their targets to spend more time and resources on information search and information processing when making their decisions, irrespective of their ultimate course of action. The aforementioned social costs of successful nudges can also translate to economic sanctions on those who resist them and thereby deviate from actual or purported social norms (e.g., Fehr and Fischbacher 2004) or simply diminish these individuals' long-term economic prospects, due to the lowering of their social image or social status (Ball et al. 2001; Bursztyn and Jensen 2017).

Finally, as with other interventions, the behavior changes produced by nudging can impose economic costs on third parties. To illustrate, HERs that lead consumers to reduce their energy consumption inevitably produce net revenue losses for energy retailers due to their diminished sales (i.e. retailers' markup above their avoidable costs). These costs can be substantial, amounting to as much as 40% of consumers' retail savings in an important natural gas conservation study (Allcott and Kessler 2019). In a similar vein, a recent reassessment of Allcott's (2011) electricity HER found the retailer net revenue losses at 25% of consumers' retail savings, at a conservative estimate (Tor and Klick 2022).

5.3 Private Opportunity Costs[7]

On many occassions, the most significant costs of most behavioral regulation are the private opportunity costs it entails for the individuals whose behavior it successfully changes. Even policies that make their targets better off on balance inevitably entail opportunity costs (OCs), due to the forgone benefits these individuals obtained from their former course of action.

5.3.1 The Private Opportunity Costs of Regulation

The inevitable imposition of private opportunity costs is most obvious for coercive regulation, which naturally forces some of its targets to modify their conduct to

7 This section is based on **Tor (2023)**.

comply with a mandate or a ban (e.g., to wear a seatbelt while driving) even when they would have been privately better off not doing so. Yet non-coercive traditional instruments, like subsidies or taxes, also entail opportunity costs. A person who decides to stop working and instead go to college because the state subsidizes her education, for example, is foregoing the benefits of the employment income now lost to her. Even traditional policies that merely provide information while changing neither their targets' incentives nor the constraints they face entail OCs for those whose behavior they change. A disclosure policy that requires the display of calorie counts on food packaging may cause some individuals to consume less of a product of whose high calorie count they previously were not fully aware. Regardless of their resulting health benefits, these people inevitably forgo the benefits they previously enjoyed from consuming more of the high-calorie product.

Some policies intentionally cause their targets to substitute personally less beneficial behaviors for more beneficial ones, thereby making such instruments privately costly on balance. This is often the case with public welfare interventions that impose net costs on some individuals to generate social benefits, such as by internalizing some negative externality they would otherwise generate (e.g., through energy consumption). Paternalistic policies are also capable of producing net private costs contrary to their stated purpose, whether due to regulator error in the face of limited information, the intentional manipulation of regulation to advance the regulators' self-interest, or the universal application of policy instruments to a heterogeneous population.

The familiar problem of honest error by regulators who cannot possess all the necessary information to guide complex economic processes—also known as the "knowledge problem" (Coase 1960; Hayek 1945)—is of particular concern for paternalistic regulation. To increase private welfare, policy makers must identify when, how, and to what extent individual judgments and decisions fall short; determine how different deviations from rationality interact both within and between individuals; find the most effective means to address these failings; and more. Hence, the complexity and scope of the necessary information increase the likelihood of error in private welfare regulation (Sunstein 2019).

In addition, the limits of human rationality revealed by behavioral research apply to regulators as well and may sometimes exacerbate the knowledge problem and other institutional challenges these decision makers face (Glaeser 2006), though regulators—who are removed from the individual choices they shape and enjoy the benefits of expert advice and deliberation—also possess certain rationality advantages (Jolls et al. 1998; Tor 2008). At any rate, regulators can make individuals worse off by mistakenly intervening when no available policy is

capable of improving people's welfare or simply by selecting the wrong instrument (e.g., banning an activity they should have taxed instead).

Other purportedly paternalistic interventions may diminish private welfare because they intentionally manipulate behavior to benefit policy makers or powerful interests they support. Public choice scholarship examines at length how decision makers within public institutions may favor personal or institutional considerations at the expense of the interests of the public they are charged to serve (Mueller 2003). In particular, bureaucrats may act to expand their power (Wilson 1989) and tend to provide inefficiently high levels of regulation (Peltzman 1976). Policy makers also can be "captured" by interest groups, such as regulated firms that have the incentives and the means to promote regulatory actions that benefit them at the expense of the diffuse public (Stigler 1971).

Finally, paternalistic regulation that uniformly applies to a heterogeneous population routinely makes some individuals worse off (Allcott and Sunstein 2015). This is because the same behavior (e.g., increased retirement savings contributions) that improves the welfare of some, even many, individuals, can be harmful to others (such as some low-income individuals who would benefit much more from using the same resources for present consumption).

Traditional paternalistic regulation that employs financial instruments, rather than mandates or bans, may be less harmful on balance but still imposes net-cost changes in behavior on some of its heterogeneous targets. This is clearly illustrated by the literature on "sin taxes," whose primary goal is to reduce individuals' consumption of some goods, such as alcohol, tobacco, or sugary drinks (O'Donoghue and Rabin 2006). When boundedly rational individuals overconsume these goods—say, because they underestimate their harmful effects or due to limited self-control (Bernheim and Taubinsky 2018)—taxes that increase their prices can fulfill a "corrective" function, leading consumers to substitute away from them. However, even sin taxes that provide society with net benefits impose net-cost behavior changes on those who did not overconsume the sin good pretax (Farhi and Gabaix 2020).

5.3.2 The Private Opportunity Costs of Behavioral Regulation

While the observation that traditional instruments can make some or all of their targets worse off on balance is well known and widely accepted, a common yet erroneous view is that the same does not hold for behavioral interventions. Thaler and Sunstein (2008), for instance, assumed that nudges' non-coercive natures guarantees they will not lead people to make privately detrimental behavior changes. This assumption is critical for the assessment of behavioral policies

because nudges that impose no net private costs would make extremely appealing regulatory instruments.

Yet non-coercive policies routinely make some individuals worse off. This is apparent where public welfare interventions are concerned, as when regulators who seek to reduce externalities nudge residential consumers to conserve electricity by sending them HERs that compare their consumption to that of their neighbors and imply the presence of a social norm favoring energy conservation (Allcott 2011). All successfully nudged households inevitably bear the opportunity costs of the forgone benefits of their previous, higher electricity usage (e.g., greater indoor comfort). Moreover, at least some of them—like those who reduce usage only to avoid the "moral tax" aspect of the nudge—bear opportunity costs that exceed their private benefits from lower energy consumption.

Other behavioral interventions to advance environmental goals clearly generate net costs for their targets. Ebeling and Lotz (2015), for instance, demonstrated the dramatic effect of default arrangements on the willingness of German consumers to choose contracts that offered more expensive energy from renewable sources over cheaper energy from non-renewable sources. The success of their default manipulation may have increased public welfare, but surely did not make better off the consumers who were nudged to pay more for their energy consumption.

Behavioral policies that advance public welfare goals beyond combating negative externalities can generate comparable effects, as illustrated by the burgeoning literature on nudging to promote prosocial behaviors, such as charitable donations. Studies in this area examine whether behavioral instruments—most notably default contribution levels, but also social norm and social comparison information, reminders, or deadlines—can increase donations (Altmann et al. 2019; Deb et al. 2014; Damgaard and Gravert 2017, 2018; Goswami and Urminsky 2016; Zarghamee et al. 2017). Regardless of their disparate effects and whether they are publicly beneficial on balance, such nudges always succeed by increasing their targets' charitable contributions at personal expense.

We noted that the only circumstances that justify paternalistic interventions are those in which people act contrary to their best interests (Bernheim and Taubinsky 2018). Like traditional instruments—e.g., sin taxes—a paternalistic behavioral intervention can modify such conduct, bringing people's behavior closer to what policy makers judge is best for them. Yet many nudges actually diminish rationality or at least exploit people's bounded rationality, in which case people's ultimate behavior reveals little about their private welfare, requiring policy makers to rely on their own judgments of whether their targets have been made better off.

This means that paternalistic nudges risk imposing net private costs, and regulators employing them must contend with the same error, manipulation, and heterogeneity challenges facing traditional interventions. In fact, the familiar possibility of regulator error due to limited information is exacerbated in the case of nudging by the twin problems of calibration and distortion. The problem of calibration concerns the design of the specific contours of a behavioral instrument to achieve its policy goal. To illustrate, regulators wishing to use a social information nudge to reduce the average consumption of high-fat foods in a target population by 20% must design its specific contours to achieve this effect. They need to decide exactly which comparison information to provide (e.g., calories vs. quantity consumed); to whom the comparison should be made (e.g., how many other consumers, selected based on which sociodemographic variables); which units would be used to describe the information provided (e.g., absolute numbers vs. percentages); how the information would be displayed (e.g., verbally or graphically, using bar charts, pie charts, or other illustrations); and more.

The implementation of most nudges involves a similar multiplicity of complex design decisions, because behavioral policies that are not well-calibrated to achieve their specific ends are likely to fail, undershoot or overshoot their mark, or even backfire (Sunstein 2018; Tor 2020a). Thus, calibration is a variant of the standard knowledge problem that poses a particular challenge for behavioral instruments.

Moreover, the very subtlety of nudging—namely, designing the boundedly rational individuals' decision environment—make it particularly difficult to calibrate. Slight changes in nudge design can produce large effects, so that seemingly comparable behavioral interventions generate very different outcomes. Consequently, a nudge can easily miss its mark, exerting too weak or—which is of particular concern with respect to opportunity costs—too strong an effect that diminishes its targets' welfare.

Ideally, regulators would engage in extensive field-testing of alternative nudge designs, in the specific context and circumstances under which they wish to adopt a behavioral policy. Such pretesting could help determine which nudges most effectively move behavior in the regulators' desired direction and, no less important, identify the specific contours that would produce the wished-for magnitude of behavior change to ensure that a paternalistically-motivated intervention does not hurt those it aims to help. To date, however, most behavioral regulation has been implemented without substantial pretesting, and even field studies that subjected some instruments to a basic empirical testing of their efficacy rarely offer sufficient evidence to enable proper nudge calibration (Allcott and Kessler 2019).

The problem of accurately calibrating paternalistic behavioral interventions is exacerbated in the case of rationality-diminishing nudges and even some

bounded-rationality exploiting nudges, which can distort people's beliefs. For example, decision makers' tendency to overestimate the likelihood of better-noted or remembered events—a phenomenon known as the availability heuristic—can be exploited by placing large, boldly-colored tickets on vehicles for parking violations, leading drivers to overestimate their probability of being ticketed to increase compliance with parking regulations (Jolls et al. 1998).

While distorting the judgments of those who might otherwise violate parking regulations—a public welfare goal—might be socially beneficial on balance, the paternalistic employment of comparable manipulation is more problematic. A case on point is the Chicago Lake Shore Drive nudge, described by Thaler and Sunstein (2008), in which policy makers distort drivers' perceptual judgments to reduce the likelihood they will suffer harm. To encourage drivers to slow down on a dangerous, repeatedly-curving stretch of the road, the city painted the lower speed limit on the road, followed by a series of white stripes. As Thaler and Sunstein (2008: 39) explain: "When the stripes first appear, they are evenly spaced, but as drivers reach the most dangerous portion of the curve, the stripes get closer together, giving the sensation that driving speed is increasing . . . . One's natural instinct is to slow down." This rationality-diminishing nudge seeks to protect drivers by distorting their speed, but it makes worse off drivers who reduce their speed despite being previously unbiased (e.g., someone arriving too late to a hospital emergency room with a life-threatening condition because the nudge led them instinctively to slow down).

Paternalistic nudges may also distort beliefs when they trigger emotions. Behavioral research shows that people often make heuristic judgments based on affective "tags" they associate with the subject of their judgment (Slovic et al. 2006). In such cases, emotional reactions—rather than cognitive assessments—may drive behavior (Loewenstein et al. 2001). Consider the possibility of encouraging employees to save more for retirement by exposing them to graphic images of retirees living in penury due to inadequate savings (say, a gentler version of the widely-used graphic warning labels on cigarette packages). This intervention could lead people to save more, but will have diminished the welfare of employees who excessively increase their retirement savings due to distorted, emotion-driven judgments.

The problem of distortion also applies to paternalistic nudges that exploit bounded rationality to shape people's decisions but incidentally impact their judgments, as in the common case of default arrangements. Researchers have identified a number of psychological processes that underlie these effects, one of which concerns the implicit recommendation embedded in policy defaults (Dinner et al. 2011; Jachimowicz et al. 2019). For instance, some individuals facing a default retirement savings rate of 6% of their salary may infer that regulators

have determined, based on information and expertise, that this rate best trades off their present consumption versus future needs, even when that is not the case.[8]

In fact, recent studies reveal that even rationality-enabling nudges that seem affect-free may carry emotional connotations that unwittingly generate costly distortions. For example, Thunström et al. (2018: 270) examined the behavioral effects of a paternalistic, money-saving reminder to consumers that stated: "Remember that the less you spend in this study, the more money you will have for other purchases." Participants who already tended to spend too little because they found spending more emotionally painful responded to the nudge by further reducing their spending, to their own detriment.

Beyond increasing their likelihood of error, the problems of calibration and distortion provide regulators further opportunities to manipulate people to their own ends. Imagine, for instance, two competing policies—one traditional, the other behavioral—that encourage people to purchase more expensive, energy-efficient, home appliances, whose expected lifetime costs are lower—and therefore more privately beneficial—than those of cheaper, less-efficient appliances. The traditional policy offers a 7% rebate on the purchase price of high-efficiency appliances, while the nudge places a highly visible "Energy Star" certification on them (cf. Houde, 2018). The problem of calibration means that it is easier to predict consumers' demand response to the 7% price reduction than to forecast their reaction to the Energy Star certification. Moreover, because the behavioral effects of Energy Star certifications will vary depending on their specific features beyond mere informational content—such as their color, size, wording, or placement—extensive testing may be necessary to identify the precise form of certification whose consumer demand effects best approximate those of the 7% rebate.

If that were not enough, the distortion problem also means that the private welfare effects of the Energy Star certification would remain ambiguous even if further testing helped calibrate the nudge. To see why, consider the reasons for which consumers increase their demand for efficient appliances following either intervention. The rebate case is straightforward—a reduction in the price of efficient appliances makes them more attractive relative to substitute products—but the welfare implications of the certification are more ambiguous, depending on the reason it changes consumer behavior. If the Energy Star merely provided

8 Some defaults exert similar effects by conveying social norm information (Davidai et al. 2012). More generally, paternalistic nudges that impact choice directly, rather than by shaping judgments, can also make their targets worse off for reasons that largely echo the preceding analysis (Tor 2020a).

consumers with information they were lacking, it could have helped them to recognize the benefits of efficient appliances and, thereby, to make better purchase decisions. However, if consumers infer from the certification that an appliance is of a higher quality rather than merely energy efficient—whether due to a misinterpretation of its meaning or because the learning of the energy-saving benefits of the certified product leads to generalized positive beliefs about the product—their distorted judgments will lead them to demand some efficient appliances whose costs exceed their private benefits.

These calibration and distortion effects also mean that regulators can more easily exploit the Energy Star to benefit themselves or relevant interest groups. For example, regulators may intentionally employ the certification to inflate consumers' beliefs in the overall quality of efficient appliances that offer manufacturers higher profit margins. Such manipulation would be far more difficult to detect and discipline than an intervention that employs an excessive price rebate to the same end.

The usual challenge of heterogeneous preferences also affects behavioral instruments much like it does traditional interventions, since nudges that encourage a uniform behavior change or a uniform ultimate behavior on the part of their targets are bound to make some of them worse off. For instance, a behavioral policy that encourages employees to save 3% of their income for retirement would be costly for all who would have been better off with a lower or higher contribution rate (cf. Choi et al. 2004)

Yet in addition to imposing private costs on some, paternalistic nudges face an additional challenge due to heterogeneity in rationality—that is, to the fact that people deviate from rationality to different degrees, depending on the specific circumstances under which they make their judgments and decisions. (Stanovich and West 1998; Tor 2014). Given heterogeneity in rationality, paternalistic nudges exert different effects on different individuals. Some will be more responsive than others to a particular behavioral intervention, at times to their personal detriment. Previously unbiased individuals may be led to make welfare-diminishing behavior changes, while formerly biased individuals may respond too strongly to a nudge. Consequently, paternalistic nudging is privately costly for some even if it produces net social benefits.

To illustrate, consider a behavioral health intervention to reduce consumption of prepared, high-fat foods by marking them with colorful hazard symbols on menus at food establishments or on the packaging of manufactured foods (e.g., Cioffi et al. 2015). For this nudge to be paternalistic, the private health risks and other costs associated with consuming such foods must exceed their nutrition, enjoyment, and other private benefits. Real consumers engage in welfare-diminishing consumption of high-fat foods for a variety of reasons: Some

may be unaware of their nutritional content; others may underestimate these products' health risks or their personal vulnerability to them; and yet additional consumers might accurately judge the risks of consuming such foods but nevertheless act myopically.

However, even if regulators make consumers collectively better off on balance—say, because the nudge draws the attention of previously inattentive consumers to the high-fat content of certain foods–they may hurt some of them due to heterogeneity in rationality. Specifically, some consumers who were already attentive to their foods' high-fat content may reduce their consumption even further because they now overestimate their health risk or experience diminished enjoyment from eating the marked foods. Alternatively, these individuals might wish to appear health conscious and thus refrain from purchasing prominently-marked foods they would have otherwise preferred to purchase. In either case, the successful paternalistic nudge will have imposed net private costs on these consumers.

5.4 Spillover Effects

In addition to their various immediate costs, nudges can also generate spillover effects by leading individuals to change other behaviors. Spillovers can be benign or even welfare increasing, such as when a net-benefit nudge facilitates other net-benefit behavior changes, but can also generate additional unintended costs (cf. Dolan and Galizzi 2015). Though potentially significant, the empirical evidence on the spillover effects of nudging is limited to date, with most scholarship considering the implications of more general evidence regarding behavioral spillover effects, which shows mixed results (Truelove et al. 2014).

Behavioral interventions that transform people's beliefs or preferences can generate positive spillovers through the very psychological mechanisms they activate. For instance, pro-environmental nudges promoting energy conservation may directly target one set of behaviors, such as those relating to household electricity use during the period of intervention, but they might impact other related behaviors as well. Indeed, studies show that the effects of HERs persist for some time after households stop receiving them—decaying at a rate of 10–20% per year (Allcott and Rogers 2014). This persistence is a positive temporal spillover (Nilsson et al. 2017), which may be due to changes in target households' beliefs about prevailing social norms or the behavior of neighbors, the development of habits (i.e., preferences) that reduce electricity consumption (Frey and Rogers 2014), or even investments in capital stock, such as efficient appliances or home improvements (Brandon et al. 2017).

However, even the most benign nudges can also produce negative spillovers, as in the case of common rationality-enabling techniques—like reminders or active choice interventions—that encourage people to pay greater attention to particular decisions or deliberate over them. Because attention is a limited resource, devoting more attention to one task inevitably depletes the amount of attention available for other tasks. Consequently, a nudge that improves performance on one task—such as identifying the healthiest main course on a menu—may diminish performance on other tasks—like that of noting the overall caloric value of the meal including side dishes and drinks (see also Altmann et al. 2021).

Negative spillovers are also likely when rationality-diminishing nudges distort beliefs, which occurred when policy makers sought to encourage public uptake of Covid-19 vaccinations or compliance with protective measures by overemphasizing or dramatizing the risks of this one disease. The successfully nudged by such means may be more likely to vaccinate or to comply with protective measures, but also more inclined to engage in privately harmful behaviors, like deferring important medical procedures or sacrificing income, social interaction, or other sources of private welfare above the level indicated by an objective risk assessment (e.g., Czeisler et al. 2020).

At other times, successful nudges can produce negative spillovers when their targets engage in other behaviors that substitute for the forgone behavior.[9] This pattern is illustrated by a recent field experiment in Brazil (Medina 2021), in which credit card holders received reminders that future payments were due. The reminders reduced average late fees but also increased overdraft fees that offset the benefits of the nudge, rendering the net effect of the intervention statistically non-significant.

Finally, a rich literature in psychology documents a number of processes that lead people who engage in one behavior to be more likely to engage in other compatible behaviors (e.g., cognitive dissonance or "foot in the door" effects) or contradictory ones (e.g., ego depletion or moral licensing), though only a handful of studies examine such processes following successful nudges (Dolan and Galizzi 2015; Truelove et al. 2014). Tiefenbeck et al. (2013) offer one such example, in a field study of water and electricity consumption. Participants targeted by a water conservation nudge used 4.1% less water, but consumed significantly more electricity, than the control, and a rough comparison suggested that the nudge's (electricity) spillover costs exceeded its (water) benefits by a factor of 2:1–6:1.

9 In principle, one can imagine situations in which the success of a nudge generates positive spillovers due to complementarities between the newly modified behavior and other behaviors (e.g., an investment in some multipurpose capital stock that enables other beneficial behaviors beyond those directly targeted by the intervention).

6 Nudge Assessment

6.1 Methods of Nudge Assessment

Ideally, a fuller appreciation of the welfare effects of nudges that includes their public and private costs as well as their benefits would enable scholars and regulators to conduct a cost-benefit analysis (CBA) of behavioral interventions, just as in the case of traditional regulation (Boardman et al. 2018; Ellig et al. 2013). CBA addresses the key economic issue with respect to policy selection—namely, identifying the most efficient option currently available to policy makers in a given context. For this reason, CBA is the dominant approach to policy assessment worldwide, mandated for U.S. federal regulation (Federal Register 1993) and playing an important role in the mandatory regulatory impact assessment processes of OECD countries (OECD 2020) and beyond (De Francesco 2012; Dunlop and Radaelli 2016).

As its name indicates, cost-benefit analysis quantifies in monetary terms the social consequences of legal interventions. While its application involves various normative challenges and technical issues, CBA's conceptual framework is straightforward: From the perspective of efficiency, the value of a policy to society is measured by its net benefits—that is, the public benefits it generates minus its public costs (Layard and Glaister 1994). Based on this assessment, CBA directs policy makers to select the option that offers the highest net benefits and to wholly avoid inefficient policies that fail to offer any net benefits vis-à-vis the status quo.

The reality of regulatory interventions, however, does not reflect CBA's *de jure* dominance. Finding that only a small fraction of federal regulation in the U.S. is assessed using the demanding methods of CBA, two researchers recently concluded that "[d]espite executive orders and office of management and budget (OMB) guidance requiring ... CBA ... of new regulations, the typical justifications and cost assessments of nonenvironmental regulations are seriously lacking" (McLaughlin and Mulligan 2020: 3). Besides the practical and conceptual challenges involved in conducting a full-fledged CBA (Boardman et al. 2018; Sunstein 2018), moreover, analysts also avoid it when they are unwilling or unable to monetize policy benefits. This is common in areas such as health or medicine, in which the monetization of benefits requires placing a monetary value on human life or quality of life that some wish to avoid (Layard and Glaister 1994).

Cost-effectiveness analysis (CEA) is the most common CBA alternative, widely employed not only in health and medicine but also in other important regulatory fields such as education (Levin and Belfield 2015), energy and the environment

(Arimura et al. 2012), and beyond (Boardman et al. 2018). Rather than calculate the monetary value of, say, the number of lives saved by the assessed intervention, CEA only monetizes policy costs, measuring benefits instead in terms of policy effectiveness vis-à-vis the status quo, using whatever metric a given policy's concrete goals offer (e.g., number of lives saved). Policy costs are then divided by effectiveness to generate a cost-effectiveness (CE) ratio that allows for a comparison of competing policies' costs per unit of effectiveness (Levin and McEwan 2001). A lower CE ratio indicates a more attractive policy that costs less per unit of effectiveness than a competing option with a higher ratio.

By retaining a broader focus that considers the costs of different interventions rather than on their effectiveness alone, CEA plays a valuable role in policy assessment. Yet its utility is limited in two crucial respects: For one, CE comparisons show which policy provides regulators with greatest "return on investment" but cannot reveal which competing policy is more efficient and may erroneously support the selection of a less-efficient policy. By beginning from the (implicit) assumption that some intervention is desirable, moreover, CEA may even favor inefficient interventions that diminish social welfare (Boardman et al. 2018; Tor and Klick 2022). In fact, CEA's real-world regulatory practice is even more problematic, as it is commonly used as part of administrative program evaluations, whose cost calculations focus on the government side of the ledger—primarily program implementation costs—to the exclusion of the often-substantial private costs these interventions generate (e.g., Allcott 2011; Ito 2015).

Furthermore, many regulatory policies are advanced without a systematic evaluation of the data necessary for even a rudimentary CEA (McLaughlin and Mulligan 2020). It is thus unsurprising that only a handful of nudge CBAs have been published to date, with CEAs being only a little more common. While researchers continue to make headway in studying their welfare effects, moreover, the assessment of behavioral policies must overcome additional obstacles. One such problem concerns the degree to which the accepted valuation methods that undergird CBA and CEA remain valid when individuals systematically deviate from the assumptions of rationality (Weimer 2017). Another challenge relates to the non-coercive nature of behavioral instruments, which leads many commentators to underestimate or even ignore the significant private costs they generate (Tor 2019). Indeed, even careful and sophisticated CBAs that grapple with the potential private costs of nudges still tend to underestimate their scope (Tor 2023).

In response to the costs and challenges of conducting a full CBA of behavioral interventions with the resulting rarity of such analyses on the one hand, and CEA's further shortcomings and remaining need for reliable cost estimates on the other, recent scholarship has proposed rationality-effects analysis (REA) as a

complementary method of nudge assessment (Tor 2019). Instead of attempting to monetize the full range of a policy's costs—a hurdle that CBA and CEA must both overcome—REA focuses on the likely effects of a nudge on the rationality of its targets. By examining these effects, which largely depend on the specific behavioral tools an intervention employs and the context within which they operate, REA can distinguish among different nudge categories that merit different treatment. For instance, rationality-promoting nudges merit a presumption allowing their implementation, while rationality-diminishing interventions bear a presumption against their adoption.

Because it does not quantify the cost and benefits of different policies, REA cannot fully substitute for CBA, or even for CEA. Yet, the assessment of nudges' rationality effects allows policy makers more quickly to determine which behavioral instruments are usually better avoided and which are more likely to produce net social benefits. The insights provided by REA are particularly valuable, moreover, when the best assessments of a nudge's benefits and costs are still highly uncertain. Additionally even when its conclusions are more equivocal, REA can identify interventions that should be prioritized for CBA, while also helping to highlight some private costs and benefits that analysts otherwise tend to ignore or underestimate.

6.2 Cost-Benefit Analysis

6.2.1 CBA of a Social Welfare Nudge: Home Energy Reports[10]

The most comprehensive CBA of a public welfare nudge currently available is Allcott and Kessler's (2019) assessment of the ubiquitous HERs. These reports' front page compares the energy use of the recipient household to that of its 100 geographically nearest neighbors in houses of a similar size, using a three-bar graph. The graph displays the household's usage against two comparison targets: one is the mean of the neighbor distribution ("All Neighbors") while the other is the 20th percentile of these neighbors ("Efficient Neighbors"). Next to the graph, the HERs' front page also displays a box that aims to signal normatively desirable behavior. Consumers with below-average usage earn one smiley face, while those below the 20th percentile earn two smiley faces. The back page provides further information about behaviors and investments that can reduce energy consumption.

10 The following discussions of CBA of public and private welfare nudges are based on **Tor (2023)** and **Tor and Klick (2022)**.

The study assessed a program that sent HERs to approximately 10,000 residential natural gas consumers in upstate New York over two heating seasons (winters), using an experimental design that allowed for the random assignment of nearly 20,000 households into either a treatment or a control group. The treatment group received standard HERs during one winter, followed by surveys that measured their willingness to pay (WTP) for another season of HERs. The effects of the reports on energy use were then measured. Allcott and Kessler (2019) also estimated the non-consumer effects of the nudge, including the socially beneficial externality reduction, the attendant net revenue loss to the utility providing the energy, and the HERs' implementation costs.

Importantly, the study found that consumers' mean WTP ($2.81) for the reports was substantially lower than their resulting savings from reducing energy expenditures ($4.91), implying that the reports imposed on consumers' additional, non-energy costs ($2.10) amounting to as much as 43% of their private benefits ($2.10/$4.91). These costs might include disutility from the social information "tax" aspect of the reports or the opportunity costs of reduced energy use (e.g., a colder home).

Yet, when including the public welfare effects of the nudge, Allcott and Kessler (2019) estimated the HERs produced an average net benefit of $0.77 per recipient, with a projected overall social value of approximately $600 million when aggregating this minute per-consumer net benefit over millions of recipients globally as of January 2017. The authors' estimates thus suggest these HERs were socially (slightly) beneficial on balance even though they imposed substantially larger net private costs on their targets, nicely illustrating the propensity of common public welfare nudges to generate such private costs.

Allcott and Kessler (2019) also found a great deal of heterogeneity in consumers' WTP for the reports. In particular, while they estimated an average seasonal net welfare gain of $0.77 per household, only 41% of these households were willing to pay more than the $1.88 marginal public cost of the nudge. This sizable minority valued the HERs highly enough to more than make up for the losses incurred by the remaining 59% of the population. Essentially, the uniform nudge functioned as a tax that may have increased overall public welfare and privately benefited a minority of its targets, but at a private net cost to their majority.

Additional unpublished evidence suggests, however, that the households' net private costs were very likely greater than the study's baseline estimate of $2.10. Allcott and Kessler (2019) report in an Online Appendix that the large majority of consumers in their study dramatically overestimated their energy savings from the HERs. This finding indicates that consumers' reported WTP was likely biased upwards and their true net private costs concomitantly greater

than the authors' baseline estimate. Given that the study's estimated $0.77 per household net social benefit, in the probable case that the WTPs' upward bias was greater than this small figure, a corrected CBA would conclude that the HERs were not only privately costly but also socially costly on balance, their public welfare goal notwithstanding.

Of further note is the dramatic difference between the outcomes of the study's more comprehensive CBA and the approach typically used to assess nudges. Specifically, studies of energy-saving nudges routinely consider implementation costs and direct energy cost savings to consumers only. Taking such an approach here would have erroneously suggested a private welfare gain of $2.69 per consumer and a public welfare gain of $1.22 billion for the HERs globally (Allcott and Kessler 2019). In other words, a failure to account for the full range of these policies' benefits and costs would have led to a two-fold overestimation of their net private and public welfare benefits alike.

6.2.2 CBA of a (Mostly) Private Welfare Nudge: Cigarette Graphic Warning Labels

Much like the case of public welfare nudges, the cost-benefit evidence concerning graphic warning labels (GWLs) on cigarette packages reveals the potentially dramatic effects of accounting for OCs in CBA. The official primary goal of GWLs is to improve individual well-being through the provision of information about the health risks associated with smoking (76 Fed. Reg. 36,627, 36,629).[11] Nevertheless, recent findings show this widespread policy involves substantial behavioral elements beyond mere information disclosure (Noar et al. 2016, 2017; Romer et al. 2018), suggesting that GWLs fit our nudge definition.

The Food and Drug Administration (FDA 2020) recently introduced its final GWL regulation, having conducted the required CBA earlier in the process. This analysis conceded that the assessed range of the monetized health benefits GWLs provide to smokers "overstate[s] . . . the net internal (i.e., intrapersonal) benefits . . . of reduced smoking because they . . . do not account for any lost consumer surplus." (FDA 2020: 36,722). According to the agency's estimates, accounting for the opportunity costs to smokers who change their behavior due to GWLs—the aforementioned "lost consumer surplus"—has a dramatic effect on the rule's CBA. The highest estimate of these OCs, which included their full monetary value despite the addictive nature of cigarettes, amounted to approximately 93% of

11 A reduction in the rate of smoking also entails substantial benefits to non-smokers, but the direct benefits and costs to smokers are central to the case for employing GWLs, as illustrated by the FDA's cost-benefit analysis (Levy et al. 2018).

the rule's benefits, even while the lowest estimate counted them as merely 10% of the same. Notably, both the FDA's analysis and most scholarship concerning the effects of tobacco control policies discount the lost benefits from smoking (i.e., their private opportunity costs) because of the addictive nature of cigarettes. Some even argue that these costs should be completely ignored, an approach that would render GWLs a far more attractive intervention than they otherwise appear (Chaloupka et al. 2015).

One need not take a stance on the appropriate discounting of the benefits to consumers from an addictive, hazardous product like cigarettes to recognize the dramatic effect of the chosen degree of discounting on how the GWL rule fares under CBA, however. As Levy and co-authors note, "the FDA analysis suggested that somewhere between almost none and almost all of the health benefits to smokers from reduced smoking are offset by lost enjoyment" (Levy et al. 2018: 5; emphasis added). This observation is especially significant for an activity such as smoking that clearly harms individuals' health, since a policy that successfully reduces its incidence will tend to generate substantial private health benefits. Nevertheless, the dramatic observation that, unless discounted, the private opportunity costs would offset nearly all of the GWLs' benefits highlights the potential impact of accounting for the OCs of private welfare nudges more generally.

Of course, GWLs are *sui generis* in important respects. For example, the addictive nature of tobacco may not only diminish the efficacy of any soft intervention in smoking behavior but also produce higher opportunity-cost estimates compared to non-addictive substances due to the increased pain of forgoing smoking. If this were case, OCs could be expected to constitute a lower fraction of the benefits of private welfare nudges that address non-addictive behaviors even while showing some resemblance to the tobacco case when targeting behaviors with an addictive component (e.g., the consumption of sugary beverages).

At the same time, the fact that most smokers do not reduce smoking after exposure to GWLs may indicate that the small minority that responds to the nudge consists mostly of those who find it easier to reduce smoking, whether because they do not enjoy it as much as other smokers do or because they are less susceptible to developing nicotine dependence (McClernon et al. 2008). Yet, if this were the case, the magnitude of the estimated opportunity costs from reduced smoking relative to its health benefits could be more comparable to those of successful private welfare nudges that target non-addictive behaviors.

Finally, even the FDA's highest OC estimate does not consider the possibility that some individuals reduce their smoking rate because the GWLs distort their beliefs or preferences, in which case they might be bearing net private costs. For instance, the labels could lead some smokers to overestimate their personal risk

of suffering the more horrific effects of smoking that are graphically displayed on the labels. GWLs could even directly diminish others' enjoyment from smoking by associating the activity with an unpleasant emotional reaction. In either case, some individuals may reduce their smoking rate beyond the level required to correct for whatever bias previously led them to smoke excessively, thereby suffering a net loss of private welfare.

Caveats of this sort are of limited concern in the case of an addictive, hazardous substance like tobacco, but similar distortionary effects would be of greater concern for paternalistic policies to shape behaviors whose harmful consequences are less pronounced or more varied—for example, in areas like retirement savings, exercise, or nutrition. In such circumstances, the various sources of increased private opportunity costs should be examined carefully to make sure that private welfare nudges in fact make net benefit policies.

6.3 Cost-Effectiveness Analysis[12]

A cost-effectiveness analysis of nudges that accounted only for their implementation costs would suggest they can be far more cost-effective than traditional instruments, as illustrated by the most notable CEA of behavioral instruments to date. This comparative CEA was conducted by Benartzi et al. (2017), a group of prominent scholars, who reviewed extant studies in major areas in which evidence concerning nudge efficacy was available at the time. In each area, they assessed the effectiveness and implementation costs of different interventions to determine the cost-effectiveness of behavioral instruments compared to that of traditional (primarily financial) interventions, repeatedly finding the best-performing nudges that they reviewed substantially more cost-effective than their best-performing traditional competitors.[13]

Nonetheless, a closer look at the studies examined by Benartzi et al. (2017) that accounts for the full range of policy costs—most notably their private opportunity costs—while excluding expenditures that are mere transfers rather than economic costs, reveals that the authors' conclusions are overstated. In reality, the very behavioral interventions they assessed were often no moreless cost-effective than effective traditional instruments.

---

12 This section is based on **Tor (2023)** and **Tor and Klick (2022)**.

13 Instead of employing the more common CE ratio that divides policy costs by a measure of its effectiveness (Benartzi et al. 2017; Levin and McEwan 2001) presented their findings using the reciprocal EC ratio, which divides effectiveness by cost. For consistency with the literature and clarity, however, their data is discussed here using the standard CE measure.

6.3.1 CEA of Public Welfare Nudges: Electricity and Gas Home Energy Reports

In what probably was the first large-scale RCT to test the effectiveness of a public welfare nudge in the area of energy conservation, Allcott (2011) studied the effects of HERs on the electrical consumption of residential homes. His basic estimate showed an average treatment effect of 2%, yielding a cost effectiveness of 3.3 cents per kilowatt-hour (kWh) saved, when dividing the administrative implementation cost of printing and sending the reports to the targeted households by the average kWh saved per year. This CE ratio was notable, comparing favorably with the estimated CE of traditional financial instruments aimed at the same goal. For example, earlier work by Arimura et al. (2012), using utility-level data from a nationwide annual panel to correlate program expenditures with changes in electricity use, produced CE estimates of approximately 6 cents/kWh, nearly double that of the HERs.

Yet, Allcott (2011) recognized that his data excludes the private costs to energy consumers, thereby understating the true cost of the HERs, further noting that if the effects of the HER nudge are due to an increased "moral cost" of consumption—that is, by making energy use psychologically or emotionally costlier—some individuals who reduce their energy consumption experience a utility loss. HERs thus entail private costs whose inclusion may reduce or eliminate their seeming cost-effectiveness advantage over traditional interventions.

This pattern is demonstrated by Allcott and Kessler's (2019) recent CBA of a natural gas HER, which provides sufficient data for a CEA of a similar nudge. The authors estimated their HERs led to an average reduction of 6.59 ccf (659 cubic feet) in natural gas use for the season they studied. Since the reports entailed an estimated implementation cost of $2.22 per HER in this case, a CEA that takes the approach used by Allcott (2011) and includes only these two effects yields a CE of $0.34 per 1 ccf reduction of energy use ($2.22/6.59 ccf). However, Allcott and Kessler (2019) were able to estimate the average non-energy costs imposed on recipient households as well as the retailers' net revenue loss (RNRL) from reduced sales, which they pegged at $2.10 and $2.53 per HER recipient. Accounting for these costs yields a corrected total cost of $6.85 ($2.22 implementation costs + $2.10 household non-energy costs + $2.53 RNRL) per HER and a CE ratio of $1.04 per ccf saved ($6.85/6.59 ccf) that is more than thrice as high as the cost-effectiveness figures generated when the nudge's private costs are ignored.

Their analysis is informative, even though it does not offer comparisons to the performance of traditional instruments in this area, as it illustrates how using appropriate cost measures—particularly the inclusion of the private costs—can render behavioral energy saving policies substantially less attractive than they

appear at first blush. If a roughly similar ratio of non-energy to energy costs were to apply to Allcott's (2011) data, for instance, the electricity HERs he studied would have exhibited CE figures similar to those estimated for the financial incentive programs evaluated by Arimura et al. (2012).

6.3.2 CEA of Private Welfare Nudges: Retirement Savings Contributions

Much like in the case of public welfare nudges, Benartzi et al. (2017) also concluded that behavioral instruments aiming to advance private welfare tend to be more cost-effective than the competing traditional financial and education-based interventions. Once again, however, this conclusion is based on a similar, erroneous, calculus that includes mere resource transfers—a particularly notable omission for the assessment of a transfer policy. Consequently, a corrected, methodologically-appropriate CEA of the competing instruments shows that nudges do not hold a consistent CE advantage over their traditional counterparts in this policy area.

On the behavioral side of the ledger, Benartzi et al.'s (2017) found a "required choice" policy studied by Carroll et al. (2009)—which merely asked employees to choose their preferred contribution rate but left them free to decide whether and how much to contribute and imposed no penalty on those who failed to make the choice—exceedingly cost effective. According to these authors' calculations, the nudge entailed only $2 of administrative costs per employee (adding a form to the company's hiring packet and making follow-up phone calls to those who did not make the choice immediately when requested to do so) while yielding a $200 average increase in annual employee contributions, thereby offering a rather astounding CE of merely $0.01 per $1 of increased contributions.

Benartzi et al. (2017) compared this intervention to five policies they classified as non-behavioral, most of which used financial incentives alone or in combination with some information provision, finding that these traditional policies exhibited CE rates ranging from $0.07 to $0.81 per $1 of increased retirement savings contributions. In other words, according to Benartzi et al. (2017), the Caroll et al. (2009) nudge was seven times more cost effective than the best-performing financial intervention of Duflo and Saez (2003) and almost two orders of magnitude more so when compared to the worst-performing instrument studied by Duflo et al. (2007).

Duflo and Saez (2003), for instance, used a field experiment at a large university, randomly selecting some employees to receive a $20 conditional voucher to participate in an employee benefit program information session. The incentive dramatically increased participation in the information session on the part of voucher recipients ("treated employees") and produced a significant relative

increase in their likelihood of joining the employer's retirement saving program (of approximately 20%) compared to an untreated control group, although this effect was small in absolute terms (a 1.25% increase from a 34% baseline). Duflo and Saez (2003) also found a statistically-indistinguishable effect on the rate of joining the retirement saving program among non-recipient employees from the academic departments in which some of their peers received vouchers ("treated departments"). Based on the estimates reported by Duflo and Saez (2003), Benartzi et al. (2017) calculated a CE of $0.07 per $1 of increased retirement savings (a total voucher cost of $12,000 divided by an estimated aggregate effect of $175,000).

However, a corrected CEA suggests a very different conclusion. Carroll et al.'s (2009) nudge indeed cost merely $2 per employee—the $200 cost it entailed for the contributing employees, who gave up the benefits of using the same resources to other ends, was a self-transfer. But Duflo and Saez's (2003) vouchers to recipients who attended the benefit fair are also transfers rather than economic costs. At the same time, this intervention still entailed some non-transfer administrative costs, to administer the vouchers and send a reminder letter to treated employees prior to the benefits fair, whose inclusion produces a corrected CE ratio of approximately of $0.01 per $1 of increased retirement savings contributions (Tor and Klick 2022), roughly comparable to that of Carroll et al.'s (2009) nudge.

A corrected CEA of the other retirement savings policies assessed by Benartzi et al. (2017) similarly shows the behaivoral and traditional instruments in this area performing at a roughly similar level, with the best-performing policy being the Danish tax instrument studied by Chetty et al. (2014).

6.4 Rationality-Effects Analysis (REA)[14]

We saw that full-fledged CBAs of behavioral policies are exceedingly rare, and even those less demanding (and less informative) CEAs are still uncommon. To help address this shortfall, Rationality Effects Analysis (REA) offers a practical complement to the costly and time-consuming quantitative cost-based methods of nudge assessment. Instead of attempting to monetize th full range of a policy's costs, REA focuses on the likely effects of a nudge on its targets' rationality, using these effects as a rough-and-ready means for assessing the desirability of behavioral instruments and the degree to which they should be prioritized for further, cost-based scrutiny (Tor 2019). In addition, the insights of REA are particularly valuable when the best assessments of a nudge's benefits and costs are still highly

14 This section is based on **Tor (2019)**.

uncertain, which diminishes the reliability of CBA for policy assessment (cf. Rizzo and Whitman 2019).

The rationality-based taxonomy of nudges (Part 2.2) classifies them into five main categories that range from techniques that are most likely to produce net welfare benefits to those that are most likely to be harmful. At the two ends of the spectrum one finds rationality-promoting instruments and their rationality-diminishing counterparts, respectively. In between these extremes, this classification distinguishes among rationality-enabling, rationality-neutral, and bounded-rationality-exploiting nudges.

Importantly, rationality effects depend not only the type of instrument employed but also, crucially, on the details of its design as well as its context and content. Many real-world policies, moreover, combine within a single intervention multiple behavioral instruments that are capable of producing different rationality effects. REA therefore examines the full range of rationality effects produced by a given nudge, with particular attention to those that are most likely to be harmful and thus merit further scrutiny.

6.4.1 REA of Public-Welfare Nudges: Home Energy Reports

HERs that aim to encourage energy conservation are perhaps the most widely-used public welfare nudge around the world or, in the colorful language of one energy consultant, "the biggest, baddest behavioral programs out there right now" (Fitzjarrald 2019). As previously discussed, Allcott (2011) was the first large-scale RCT to study the effects of HERs on residential electrical consumption, while Allcott and Kessler (2019) examined the welfare effects of a similar program aimed at natural gas consumption.

HERs utilize both social and non-social information nudging that are relevant for their REA. On the social information side, the reports' front pages compare the target household's energy consumption to that of select peers. REA recognizes that this aspect of the HERs can produce a number of rationality effects: Most benignly, it provides pure social information that lets people learn how they compare to an average neighbor with a similarly-sized home, making it mostly a rationality-neutral nudge, which neither helps recipients better manifest their energy consumption preferences nor hinders them from doing so. Yet this information may indicate to certain recipients—such as those who find that they consume much more or much less than comparable neighbors—that they may have overlooked ways to benefit from reducing or increasing their energy use. In such cases, the social information might even qualify as a rationality-enabling nudge that helps some households better align their behavior with their preferences.

At the same time, however, the comparative aspect of the reports intentionally highlights the performance of the most efficient 20% of homes, in the hope of activating recipients' social comparison concerns. When successful, the activation of such concerns can lead households towards greater energy conservation efforts to outdo their neighbors, or at least to avoid being outperformed by them (e.g., Garcia et al. 2013, 2020; Garcia and Tor 2022), while a failure to accomplish either of these goals can impose on the many higher-consumption households the psychological cost of an ongoing upward social comparison (Suls and Wheeler 2000). Hence, unlike more general information about the overall distribution, the HERs' emphasis on this comparison target is a bounded rationality exploiting nudge that sets high performers as a relevant comparison to shape the behavior of their peers towards the increased energy conservation favored by regulators.

The social nudging aspects of the reports go even further, however, in prominently displaying a box that seeks to construct an injunctive norm favoring lower energy consumption. This front-page feature emphasizes the recipient household's performance on a 3-level scale accompanied by emoticons. For example, the scale might range from "great" (commensurate with the top 20%) to "good" (consuming less than average) to "using more than average", accompanied by a broadly smiling face, a slightly smiling face, and a frowning face that was later dropped from the HERs, respectively. This normative box goes beyond the exploitation of recipients' bounded rationality, misleadingly implying the existence of an injunctive social norm of lower energy consumption that justifies social disapprobation towards those who consume the average amount of energy or more. Of course, below-average consumption is unlikely to reflect an extant norm, and the top 20% of the group even less so. Yet recipients who erroneously believe their energy consumption violates a social norm may suffer direct psychological costs and even change their behavior in an effort to comply with this norm (Nolan et al. 2008). Moreover, even consumers who are not misled by the normative messaging may mistakenly infer from it that a personal sacrifice on their part would produce substantial social benefits in reducing the negative externalities of energy consumption, while in reality such efforts appear to produce very small net benefits, if any (e.g., Allcott and Kessler 2019). Hence, the normative box aspect of the HER amounts to a rationality-diminishing nudge that is capable of diminishing the welfare of its targets and perhaps even social welfare overall.

Finally, the back pages of the reports also provide recipients with non-social information, including tips regarding household behaviors (e.g., turning off lights in unoccupied rooms) and low-cost home-improvement investments (such as weather-stripping external doors) that can reduce energy consumption. To the extent that these tips either remind people of behaviors they wished but

forgot to perform or inform them of effective ways to manifest their energy saving preferences, they constitute beneficial rationality-enabling element of this instrument.

Nonetheless, even these aspects of the reports can prove problematic, depending on the specifics of their design. For one, even simple reminders, particularly when they convey urgency, can pressure individuals to act contrary to their own preferences or to make decisions (e.g., regarding an energy saving investment) they are unsure about. More troublingly, however, to encourage conservation, the HERs emphasize the most optimistic cost-saving outcomes of the behaviors they promote (i.e., "save up to $100 a year"), instead of providing more realistic, representative, information (e.g., average savings from the behavior). When households act to conserve energy because they were led to overestimate their expected private benefits from onerous or financially costly behaviors, the non-social tips or reminders will have turned from a rationality-enabling intervention into a rationality-diminishing one.

The REA of Home Energy Reports thus paints a complex picture. To promote energy conservation, these interventions draw on a number of behavioral and informational instruments with widely differing rationality effects. In principle, the non-social aspects of the reports and even the basic social information they provide about peers' energy use can be designed as rationality-enabling or at least rationality-neutral policies. Yet, in practice, even these potentially beneficial HER elements are designed to tip the scales in favor of increasing energy conservation behaviors in conjunction with an effort to exploit recipients' bounded rationality by recruiting and even facilitating their social comparison concerns. If this were not enough, the reports further strive to construct novel injunctive norms and express disapprobation of recipients who fail to meet these norms, thereby risking distorting consumers' beliefs regarding the nature of prevailing norms. These potential distortionary effects are further reinforced by the HERs' exaggerated suggestions—both explicit and implicit—concerning the magnitude of the public and private benefits that would follow recipients' energy conservation behavior. As implemented, therefore, the reports at best make a bounded rationality exploiting nudge, while at worst they risk diminishing the rationality of their recipients.

This conclusion, which places HERs as implemented athwart the bounded rationality exploiting and rationality diminishing categories, suggests that the reports risk diminishing the welfare of many recipients by imposing on them direct costs or by encouraging net-cost behavior changes on their part. Indeed, the presence of these private HER costs is further corroborated by Allcott and Kessler's (2019) careful welfare analysis of a large set of natural gas HERs.

Nevertheless, REA does not hold that such behavioral interventions cannot be beneficial on balance, since they are still capable of produce net benefits for some of their recipients and of generating some social benefits, despite their rationality-related shortcomings. The latter possibility is particularly important in the case of a public welfare intervention to reduce negative externalities, since sufficiently enough benefits from such a reduction may outweigh the net private costs the nudge risks imposing on many of its targets.

Faced with the substantial conflicting effects of the HERs are currently designed, therefore, REA does not condone their implementation without a full CBA.[15] This conclusion stands in obvious contrast to the widespread adoption of HERs based on the belief that they offer a cost-effective means for reducing household energy consumption.

6.4.2 REA of Private-Welfare Nudges: Retirement Savings

The required active choice intervention studied by Carroll et al. (2009) illustrates a private welfare nudge to increase retirement savings. This nudge asked newly-hired employees at a Fortune 500 company in the financial-services sector to choose their preferred retirement savings contribution rate within 30 days of their hiring, but left them free to decide whether to join the plan and how much to contribute (imposing no penalty on those who failed to make the choice within the required time frame).

An examination of the rationality effects of this active choice nudge suggests that it qualifies as a rationality-enabling policy. In the main, the intervention encourages employees to deliberate and manifest their preferences regarding the difficult tradeoff between increased retirement savings and current consumption (or other forms of savings). This conclusion holds even though the nudge—which was implemented by adding another form to an already substantial amount of hiring paperwork and required employees to make an important financial decision with potential long-term implications as soon as they started a new job—clearly imposed some direct decision costs on its targets.

REA presumes that the adoption of policies that tend to be rationality-enabling should be permitted when these policies are efficacious, unless the specific details of their design, context, or content suggest the need for further scrutiny. For example, while neutral active choice nudges such as those studied by Carroll et al. (2009) are usually benign, active choice interventions are highly

15 Interestingly, as shown by our analysis of Allcott and Kessler's (2019) findings including their online appendix data, it is doubtful whether the HERs in fact make net benefit interventions (cf. Bernheim and Taubinsky 2018).

suspect when they incorporate additional bounded rationality exploiting or even rationality-diminishing elements, like the designation of some options as defaults or the use of language that aims to trigger loss aversion (Keller et al. 2011).

These illustrative REAs of public and private welfare nudges thus demonstrate both the benefits and the limitations of this approach to behavioral policy assessment. Most significantly, the assessment of rationality effects allows policy makers more quickly to determine which efficacious nudges are usually better avoided and which are more likely to produce net social benefits, and can even help guide the design of interventions that draw on multiple behavioral instruments simultaneously, as is often the case. Furthermore, even when its conclusions are equivocal, REA serves the important function of identifying which interventions should be prioritized for CBA, while also helping to highlight significant private costs and benefits of behavioral regulations that analysts otherwise tend to ignore or underestimate. At the same time, REA cannot fully substitute for CBA or CEA (which require a quantification of policy costs), nor can it tell policy makers whether a given intervention is likely to be efficacious—a question that necessitates an empirical investigation.

7 Conclusion

By now, behavioral regulation increasingly pervades all major areas of public policy as both a substitute for and a complement to traditional regulation. After reviewing the development of nudge definitions, part 1 explained what renders some definitions more useful than others and argued in favor of considering as nudges only those *significantly behavioral instruments*—that is, policies whose impact is due in significant part to the activation of behavioral processes that rational actors would find irrelevant. Following this delineation of the outer boundaries of nudging, part 2 turned to the internal organization of these behavioral instruments. It reviewed the main types of extant nudge taxonomies and evaluated their advantages and disadvantages, based on which it articulated a new, welfare-relevant taxonomy that differentiates among nudge techniques based on their rationality effects.

Part 3 then describes the main reasons for the increasing employment of behavioral regulation around the globe, focusing on the reasons for which nudges appeal to regulators, as well as the developing evidence for efficacy of these instruments in changing behavior. Part 4 considered the myriad public and private costs of behavioral regulation. While examining the government implementation costs of nudges and their direct private costs and spillover costs, this part emphasized

the private opportunity costs of successful nudges, which often represent the greatest costs of these interventions.

Part 5 then drew on the preceding parts and recent empirical evidence to demonstrate the CBA, CEA, and REA of both public and private welfare nudges, in each case noting the benefits and limitations of these different methods of policy assessment. The clear overall impression from a careful analysis of behavioral regulation is that while such policies can produce substantial benefits in specific cases, they also carry with them significant risks and costs that analysts typically ignore or underestimate. Regulators will therefore be well-served by a more cautious and considered approach to nudging, which not only tests the efficacy of specific interventions, but also routinely subjects them at least to a careful REA, if not a full CBA, prior to their adoption.

References

Allcott, H. (2011). Social norms and energy conservation. J. Publ. Econ. 95: 1082–1095.
Allcott, H. and Sunstein, C. (2015). Regulating internalities. J. Pol. Anal. Manag. 34: 698–705.
Allcott, H. and Kessler, J.B. (2019). The welfare effects of nudges: a case study of energy use social comparisons. Am. Econ. J. Appl. Econ. 11: 236–276.
Allcott, H. and Rogers, T. (2014). The short-run and long-run effects of behavioural interventions: experimental evidence from energy conservation. Am. Econ. Rev. 104: 3003–3037.
Allcott, H. and Taubinsky, D. (2015). Evaluating behaviorally motivated policy: experimental evidence from the lightbulb market. Am. Econ. Rev. 105: 2501–2538.
Altmann, S., Falk, A., Heidhues, P., Jayaraman, R., and Teirlinck, M. (2019). Defaults and donations: evidence from a field experiment. Rev. Econ. Stat. 101: 808–826.
Altmann, S., Grunewald, A., and Radbruch, J. (2021). Interventions and cognitive spillovers. Rev. Econ. Stud. 1–36 (in press).
Arad, A. and Rubinstein, A. (2018). The people's perspective on libertarian-paternalistic policies. J. Law Econ. 61: 311–333.
Arimura, T.H., Li, S., Newell, R.G., and Palmer, K. (2012). Cost-effectiveness of electricity energy efficiency programs. Energy J. 33: 63–99.
Arno, A. and Thomas, S. (2016). The efficacy of nudge theory strategies in influencing adult dietary behaviour: a systemic review and meta-analysis. BMC Publ. Health 16: 1–11.
Baldwin, R. (2014). From regulation to behaviour change: giving nudge the third degree. Mod. Law Rev. 77: 831–857.
Ball, S., Eckel, C., Grossman, Phillip J., and Zame, W. (2001). Status in markets. Q. J. Econ. 116: 161–188.
Bang, H., Shu, S., and Weber, E. (2018). The role of perceived effectiveness on the acceptability of choice architecture. Behav. Public Policy 4: 50–70.
Barton, A. and Grune-Yanoff, T. (2015). From libertarian paternalism to nudging—and beyond. Rev. Philos. Psychol. 6: 341–359.
Bauer, J.M. and Reisch, L.A. (2019). Behavioural insights and (Un)healthy dietary choices: a review of current evidence. J. Consum. Pol. 42: 3–45.

Behavioral Insights Team (2012). Annual update 2011–2012, Available at: https://www.gov.uk/government/publications/behavioural-insights-team-annual-update.

Behavioral Insights Team (2019). Annual report 2017–2018, Available at: https://www.bi.team/wp-content/uploads/2019/01/Annual-update-report-BIT-2017-2018.pdf.

Benartzi, S., Beshears, J., Milkman, K., Cass, S., Thaler, R., Shankar, M., Tucker-Ray, W., Congdon, W., and Galing, S. (2017). Should governments invest more in nudging? Psychol. Sci. 28: 1041–1055.

Bernheim, B.D. and Taubinsky, D. (2018). Behavioral public economics – foundations and applications 1. In: Bernheim, B.D., DellaVigna, S., and Libson, D. (Eds.), *Handbook of behavioral economics*. North Holland: Elsevier.

Berthet, V. and Ouvard, B. (2019). Nudge: towards a consensus view? Psychol. Cognitive Sci. 5: 1–5.

Boardman, A.E., Greenberg, D.H., Vining, A.R., and Weimer, D.L. (2018). *Cost-benefit analysis: concepts and practice*, 5th ed Cambridge University Press, Cambridge, UK.

Brandon, A., Ferraro, Paul J., List, John A., Metcalfe, Robert D., Price, Michael K., and Rundhammer, Florian. (2017). Do the effects of social nudges persist? theory and evidence from 38 natural field experiments. In: *NBER Working Paper #23277*. Available at: https://www.nber.org/papers/w23277.

Bursztyn, L. and Jensen, R. (2017). Social image and economic behavior in the field: identifying, understanding, and shaping social pressure. Annu. Rev. Econom. 9: 131–153.

Byerly, H., Balmford, A., Ferraro, P.J., Wagner, C., Palchak, E., Polasky, S., Ricketts, T.H., Schwartz, A.J., and Fisher, B. (2018). Nudging pro-environmental behavior: evidence and opportunities. Front. Ecol. Environ. 16: 159–168.

Capraro, V., Jagfeld, G., Klein, R., Mul, M., and van de Pol, I. (2019). Increasing altruistic and cooperative behaviour with simple moral nudges. Sci. Rep. 9: 1880.

Carroll, G.D., Choi, J.J., Laibson, D.I., Madrian, B., and Metrick, A. (2009). Optimal defaults and active decisions. Q. J. Econ. 124: 1639–1674.

Chaloupka, F.J., Gruber, J., and Warner, K.E. (2015). Accounting for 'lost pleasure' in a cost–benefit analysis of government regulation: the case of the food and drug administration's proposed cigarette labeling regulation. Ann. Intern. Med. 162: 64–66.

Chetty, R., Friedman, J.N., Leth-Petersen, S., Nielsen, T.H., and Olsen, T. (2014). Active vs. passive decisions and crowd-out in retirement savings accounts: evidence from Denmark. Q. J. Econ. 129: 1141–1219.

Choi, James J., Laibson, D., Madrian, Brigitte C., Metrick, A., and Poterba, James M. (2004). For better or for worse: default effects and 401(k) savings behavior. In: Wise, David A. (Ed.), *Perspectives on the economics of aging*. Chicago: University of Chicago Press.

Cioffi, Catherine E., Levitsky, David A., Pacanowski, Carly R., and Bertz, F. (2015). A nudge in a healthy direction. The effect of nutrition labels on food purchasing behaviors in university dining facilities. Appetite 92: 7–14.

Coase, R.H. (1960). The problem of social cost. J. Law Econ. 3: 1–44.

Czeisler, Mark E., Marynak, K., Clarke, Kristie E.N., Salah, Z., Shakya, I., Thierry, JoAnn M., Ali, N., McMillan, H., Wiley, Joshua F., Weaver, Matthew D., et al. (2020). Delay or avoidance of medical care because of COVID-19-related concerns. MMWR (Morb. Mortal. Wkly. Rep.) 69: 1250–1257.

Damgaard, Mette T. and Gravert, C. (2017). Now or never! The effect of deadlines on charitable giving. J. Behav. Exp. Econ. 66: 78–87.

Damgaard, Mette T. and Gravert, C. (2018). The hidden costs of nudging: experimental evidence from reminders in fundraising. J. Publ. Econ. 157: 15–26.
Davidai, S., Gilovich, T., and Lee, R.D. (2012). The meaning of default options for potential organ donors. Proc. Natl. Acad. Sci. U.S.A. 109: 15201–15202.
Deb, R., Gazzale, R.S., and Kotchen, M.J. (2014). Testing motives for charitable giving: a revealed-preference methodology with experimental evidence. J. Publ. Econ. 120: 181–192.
De Francesco, F. (2012). Diffusion of regulatory impact analysis among OECD and EU member states. Comp. Polit. Stud. 45: 1277–1305.
De Haan, T. and Linde, J. (2018). 'Good nudge lullaby': choice architecture and default bias reinforcement. Econ. J. 128: 1180–1206.
DellaVigna, S. and Linos, E. (2020). RCTs to scale: comprehensive evidence from two nudge units. In: *SSRN working paper*, Available at: https://www.nber.org/papers/w27594.
Dinner, I., Johnson, E.J., Goldstein, D.G., and Liu, K. (2011). Partitioning default effects: why people choose not to choose. J. Exp. Psychol. Appl. 17: 332–341.
Dolan, P. and Galizzi, Matteo M. (2015). Like ripples on a pond: behavioral spillovers and their implications for research and policy. J. Econ. Psychol. 47: 1–16.
Duflo, E. and Saez, E. (2003). The role of information and social interactions in retirement plan decisions: evidence from a randomized experiment. Q. J. Econ. 118: 815–842.
Duflo, E., Gale, W., Liebman, J., Peter, O., and Saez, E. (2007). Savings incentives for low- and moderate-income families in the United States: why is the saver's credit not more effective? J. Eur. Econ. Assoc. 5: 647–661.
Dunlop, C.A. and Radaelli, C.M. (2016). *Handbook of regulatory impact assessment*. Edward Elgar Publishing, London, United Kingdom.
Ebeling, F. and Lotz, S. (2015). Domestic uptake of green energy promoted by opt-out tarrifs. Nat. Clim. Change 5: 868–871.
Ellig, J., McLaughlin, P.A., and Morrall, J. (2013). Continuity, change, and priorities: the quality and use of regulatory analysis across US administrations. Regul. Govern. 7: 153–173.
European Commission (2016). *Behavioral insights applied to policy: European report 2016*, Available at: https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/behavioural-insights-applied-policy-european-report-2016.
Exec. order no. 12,866, 58 fed. reg. 51,735 (1993), Available at: https://www.archives.gov/files/federal-register/executive-orders/pdf/12866.pdf (Accessed 4 October 1993).
Fehr, E. and Fischbacher, U. (2004). Third-party punishment and social norms. Evol. Hum. Behav. 25: 63–87.
Farhi, E. and Gabaix, X. (2020). Optimal taxation with behavioral agents. Am. Econ. Rev. 110: 298–336.
Fitzjarrald, B. (2019). Utilities investing in behavior? Yes! Examples of behavior strategies in action. In: *Behavior, energy, and climate change webinar series*, Available at: https://www.youtube.com/watch?v=c0w_rZSaEmw (Accessed 17 November 2019).
Food and Drug Administration (2020). Tobacco products; required warnings for cigarette packages and advertisements, final rule. 85 Federal Register 15638, pp. 15638–15710, 21 CFR 1141.
Forberger, S., Resich, L., Kampfmann, T., and Zeen, H. (2019). Nudging to move: a scoping review of the use of choice architecture interventions to promote physical activity in the general population. Int. J. Behav. Nutr. Phys. Activ. 16: 1–14.

Frey, E. and Rogers, T. (2014). Persistence: how treatment effects persist after interventions stop. Policy Insights Behav. Brain Sci. 1: 172–179.
Garcia, S.M. and Tor, A. (2022). *Social comparison and competition: a progress report*. In: Garcia, S.M., Tor, A., and Elliott, A.S. (Eds.), *Oxford handbook on the psychology of competition*. Oxford University Press, Oxford, United Kingdom.
Garcia, Stephen M., Tor, A., and Schiff, Tyrone M. (2013). The psychology of competition: a social comparison perspective. Perspect. Psychol. Sci. 8: 634–650.
Garcia, Stephen M., Reese, Zachary A., and Tor, A. (2020). Social comparison before, during, and after the competition. In: Suls, J., Collins, R., and Wheeler, L. (Eds.), *Social comparison, judgment and behavior*. Oxford: Oxford University Press.
Glaeser, Edward L. (2006). Paternalism and psychology. Univ. Chic. Law Rev. 73: 133–156.
Gold, N., Lin, Y., Ashcroft, R., and Osman, M. (2020). 'Better off, as judged by themselves': do people support nudges as a method to change their own behavior? Behav. Public Policy: 1–30.
Goswami, I. and Urminsky, O. (2016). When should the ask be a nudge? The effect of default amounts on charitable giving. J. Market. Res. 53: 829–846.
Grune-Yanoff, T. and Hertwig, R. (2016). Nudges versus boost: how coherent. Minds Mach. 26: 149–183.
Hagmann, D., Ho, E.H., and George, L. (2019). Nudging out support for a carbon tax. Nat. Clim. Change 9: 484–489.
Hall, J.D. and Madsen, J. (2021). Can behavioral interventions be too salient? Evidence from traffic safety messages. In: SSRN working paper, Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3633014.
Halpern, D. (2015). *Inside the nudge unit*. London, UK: WH Allen.
Hayek, F.A. (1945). The use of knowledge in society. Am. Econ. Rev. 35: 519–530.
Hollands, G., Bignardi, G., Johnston, M., Kelly, M.P., Ogilvie, D., Petticrew, M., Prestwich, A., Shemilt, I., Sutton, S., and Marteau, T.M. (2017). The TIPPME intervention typology for changing environments to change behaviour. Nat. Human Behav. 1: 1–9.
Houde, S. (2018). How consumers respond to product certification and the value of energy information. Rand J. Econ. 49: 453–477.
Hummel, D. and Maedche, A. (2019). How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. J. Behav. Exp. Econ. 80: 47–58.
Ito, K. (2015). Asymmetric incentives in subsidies: evidence from a large-scale electricity rebate program. Am. Econ. J. Econ. Pol. 7: 209–237.
Jachimowicz, J.M., Duncan, S., Weber, E.U., and Johnson, E.J. (2019). When and why defaults influence decisions: a meta-analysis of default effects. Behav. Public Policy 3: 159–186.
Janis, I.L. and Mann, L. (1977). *Decision making: a psychological analysis of conflict, choice, and commitment*. Free Press, New York.
Johnson, E.J., Shu, S.B., Dellaert, B.G.C., Fox, C., Goldstein, D.G., Häubl, G., Larrick, R.P., Payne, J.W., Peters, E., Schkade, D., et al. (2012). Beyond nudges: tool of a choice architecture. Market. Lett. 23: 487–504.
Jolls, C., Sunstein, Cass R., and Thaler, R. (1998). A behavioral approach to law and economics. Stanford Law Rev. 50: 1471–1550.
Jones, R., Pykett, J., and Whitehead, M.J. (2013). *Changing behaviours: the rise of the psychological state*. Edward Elgar Publishing, Cheltenham, UK.
Jung, J. and Mellers, B. (2016). American attitudes toward nudges. Judgm. Decis. Mak. 11: 62–74.

Keller, Punam A., Harlam, B., George, L., and Volpp, K.G. (2011). Enhanced active choice: a new method to motivate behavior change. J. Consum. Psychol. 21: 376–383.
Khern-am-nuai, W., Yang, W., and Li, N. (2017). Using context-based password strength meter to nudge users' password generating behavior: a randomized experiment. In: *Proceedings of the 50th hawaii international conference on system sciences*, pp. 587–596.
Klick, J. and Mitchell, G. (2006). Government regulation of irrationality: moral and cognitive hazards. Minn. Law Rev. 90: 1620–1663.
Layard, R. and Glaister, S. (1994). *Cost-benefit analysis*, 2nd ed. Cambridge University Press, Cambridge, UK.
Le Grand, J. and New, B. (2015). *Government paternalism: nanny state or helpful friend?* Princeton University Press, Princeton, NJ.
Legros, S. and Cislaghi, B. (2020). Mapping the social-norms literature: an overview of reviews. Perspect. Psychol. Sci. 15: 62–80.
Levin, H.M. and Belfield, C. (2015). Guiding the development and use of cost-effectiveness analysis in education. J. Res. Educ. Eff. 8: 400–418.
Levin, H.M. and McEwan, P.J. (2001). *Cost-effectiveness analysis*, 2nd ed. Sage Publications, Thousand Oaks, CA.
Levy, H., Norton, E.C., and Smith, J.A. (2018). Tobacco regulation and cost-benefit analysis: how should we value foregone consumer surplus? Am. J. Health Econ. 4: 1–25.
Lin, Y., Osman, M., and Ashcroft, R. (2017). Nudge: concept effectiveness and ethics. Basic Appl. Soc. Psychol. 39: 1–14.
Loewenstein, G.F., Hsee, C.K., Weber, E.U., and Welch, N. (2001). Risk as feelings. Psychol. Bull. 127: 267–286.
Madrian, Brigitte C. (2014). Applying insights from behavioral economics to policy design. Annu. Rev. Econom. 6: 663–688.
Mathis, K. and Tor, A. (2016). *Nudging – possibilities, limitations and applications in European law and economics*. Springer International Publishing, Switzerland.
McClernon, Francis J., Koznik, Rachel V., and Rose, Jed E. (2008). Individual differences in nicotine dependence, withdrawal symptoms, and sex predict transient fMRI-BOLD responses to smoking cues. Neuropsychopharmacology 33: 2148–2157.
McLaughlin, P.A. and Mulligan, C.B. (2020). Three myths about federal regulation. In: *NBER working paper #27233*, Available at: https://www.nber.org/papers/w27233.
Medina, Paolina C. (2021). Side effects of nudging: evidence from a randomized intervention in the credit card market. Rev. Financ. Stud. 34: 2580–2607.
Michie, S., Johnston, M., Francis, J., Hardeman, W., and Eccles, M. (2008). From theory to intervention: mapping theoretically derived behavioural determinants to behavior change techniques. Appl. Psychol. Int. Rev. 57: 660–680.
Mollenkamp, M., Zeppernick, M., and Jonas, S. (2019). The effectiveness of nudges in improving the self-management of patients with chronic diseases: a systematic literature review. Health Pol. 123: 1199–1209.
Morris, Michael W., Hong, Y., Chiu, C., and Liu, Z. (2015). Normology: integrating insights about social norms to understand cultural dynamics. Organ. Behav. Hum. Decis. Process. 129: 1–13.
Mueller, D.C. (2003). *Public choice III*. Cambridge University Press, Cambridge, UK.
Münscher, R., Vetter, M., and Scheuerle, T. (2015). A review and taxonomy of choice architecture techniques. J. Behav. Decis. Making 29: 511–524.

Nilsson, A., Bergquist, M., and Schultz, Wesley P. (2017). Spillover effects in environmental behaviors, across time and context: a review and research agenda. Environ. Educ. Res. 23: 573–589.

Noar, S.M., Hall, M.G., Francis, D.B., Ribisl, K.M., Pepper, J.K., and Brewer, N.T. (2016). Pictorial cigarette pack warnings: a meta-analysis of experimental studies. Tobac. Control 25: 341–354.

Noar, S.M., Francis, D.B., Bridges, C., Sontag, J., Brewer, N.T., and Ribisl, K.M. (2017). Effects of strengthening cigarette pack warnings on attention and message processing: a systematic review. Journal. Mass Commun. Q. 94: 416–442.

Nolan, J.M., Schultz, W., Cialdini, R., Goldstein, N.J., and Griskevicius, V. (2008). Normative social influence is underdetected. Pers. Soc. Psychol. Bull. 34: 913–923.

O'Donoghue, T. and Rabin, M. (2006). Optimal sin taxes. J. Publ. Econ. 90: 1825–1849.

Oliver, A. (2015). Nudging, shoving, and budging: behavioral-economic informed policy. Publ. Adm. 93: 700–714.

Oliver, A. (2017). *The Origins of behavioural public policy*. Cambridge University Press, Cambridge, UK.

Organisation for Economic Co-operation and Development (2017). *Behavioral insights and public policy: lessons from around the world*. OECD Publishing, Paris, FR, Available at: https://read.oecd-ilibrary.org/governance/behavioural-insights-and-public-policy_9789264270480-en.

Organisation for Economic Co-operation and Development (2020). *Regulatory policy and COVID-19: behavioural insights for fast-paced decision making*. OECD Publishing, Paris, DR. Available at: https://read.oecd-ilibrary.org/view/?ref=137_137590-2p5x0tveyp&title=Regulatory-policy-and-COVID-19-Behavioural-insights-for-fast-paced-decision-making&_ga=2.85898713.1392342847.1628886762-1064373644.1624924646.

Peltzman, S. (1976). Toward a more general theory of regulation. J. Law Econ. 19: 211–240.

Reisch, L. and Sunstein, Cass R. (2016). Do Europeans like nudges? Judgm. Decis. Mak. 11: 310–325.

Required warnings for cigarette packages and advertisements, 76 fed. reg. 36,627, 36,629 (September 22, 2011) (codified at 21 C.F.R. pt. 1141).

Rizzo, M. and Whitman, G. (2019). *Escaping paternalism: rationality, behavioral economics, and public policy* (Cambridge studies in economics, choice, and society). Cambridge University Press, Cambridge.

Romer, D., Ferguson, Stuart G., Strasser, Andrew A., Evans, Abigail T., Tompkins, M.K., Macisco, J., Fardal, M., Tusler, M., and Peters, E. (2018). Effects of pictorial warning labels for cigarettes and quit-efficacy on emotional responses, smoking satisfaction, and cigarette consumption. Ann. Behav. Med. 52: 53–64.

Sibony, A. and Alemanno, A. (2015). The emergence of behavioural policy-making: a European perspective. In: Alemanno, A. and Sibony, A. (Eds.), *Nudge and the law: a European perspective*. Hart Publishing, Oxford, UK.

Slovic, P., Finucane, M., Peters, E., and MacGregor, D.G. (2006). The affect heuristic. In: Lichtenstein, S. and Slovic, P. (Eds.), *The construction of preference*. Cambridge University Press, Cambridge, UK.

Spiegler, R. (2015). On the equilibrium effects of nudging. J. Leg. Stud. 44: 389–416.

Stanovich, Keith E. and West, Richard F. (1998). Individual differences in framing and conjunction effects. Think. Reas. 4: 289–317.

Stigler, G.J. (1971). The theory of economic regulation. Bell J. Econ. Manag. Sci. 2: 3–21.
Suls, J. and Wheeler, L. (2000). *Handbook of social comparison: theory and research*. Springer, New York.
Sunstein, C.R. (2014). Choosing not to choose. Duke Law J. 64: 1–52.
Sunstein, C.R. (2015). The ethics of nudging. Yale J. Regul. 32: 413–450.
Sunstein, C.R. (2016). The council of psychological advisers. Annu. Rev. Psychol. 67: 713–737.
Sunstein, C.R. (2018). *The cost-benefit revolution*. MIT Press, Cambridge, MA.
Sunstein, C.R. (2019). Ruining popcorn? The welfare effects of information. J. Risk Uncertain. 58: 121–142.
Sunstein, C.R. and Reisch, L.A. (2019). *Trusting nudges: toward a bill of rights for nudging*, 1st ed Routledge, Abingdon, UK.
Sunstein, C.R. and Thaler, R.H. (2003). Libertarian Paternalism is Not an oxymoron, 70. Univ. Chic. Law Rev., 1159–1202.
Sunstein, C.R., Reisch, L.A., and Kaiser, M. (2019). Trusting nudges? Lessons from an international survey. J. Eur. Publ. Pol. 26: 1417–1443.
Szaszi, B., Palinkas, A., Palfi, B., Szollosi, A., and Aczel, B. (2018). A systematic scoping review of the choice architecture movement: toward understanding when and why nudges work. J. Behav. Decis. Making 31: 355–366.
Teichman, D. and Underhill, K. (2021). Infected by bias: behavioral science and the legal response to COVID-19. Am. J. Law Med. 47: 205–248.
Thaler, R. and Sunstein, Cass R. (2008). *Nudge: improving decisions about health, wealth, and happiness*. New York, NY: Penguin Books.
Thorgerson, J. and Olander, F. (2003). Spillover of environment-friendly consumer behavior. J. Environ. Psychol. 23: 225–236.
Thunström, L. (2019). Welfare effects of nudges: the emotional tax of calorie menu labeling. Judgm. Decis. Mak. 14: 11–25.
Thunström, L., Gilbert, B., and Jones Ritten, C. (2018). Nudges that hurt those already hurting—distributional and unintended effects of salience nudges. J. Econ. Behav. Organ. 153: 267–282.
Tiefenbeck, V., Staake, T., Roth, K., and Sachs, O. (2013). For better or for worse? Empirical evidence of moral licensing in a behavioral energy conservation campaign. Energy Pol. 57: 160–171.
Tor, A. (2008). The methodology of the behavioral analysis of law. Haifa Law Rev. 4: 237–327.
Tor, A. (2014). Understanding behavioral antitrust. Tex. Law Rev. 92: 573–667.
Tor, A. (2016). The critical and problematic role of bounded rationality in nudging. In: Mathis, K. and Tor, A. (Eds.), *Nudging – possibilities, limitations, and applications in European law and economics*. Springer, Cham, Switzerland.
Tor, A. (2019). All nudges are not the same: why rationality matters for welfare, unpublished manuscript.
Tor, A. (2020a). Nudges that should fail? Behav. Public Policy 4: 316–342.
Tor, A. (2020b). The target opportunity costs of successful nudges. In: Mathis, K. and Tor, A. (Eds.), *Consumer law and economics*. Springer, Cham, Switzerland.
Tor, A. (2021a). A better nudge definition, unpublished manuscript.
Tor, A. (2021b). Organizing the behavioral toolbox: a rationality-based nudge taxonomy, unpublished manuscript.
Tor, A. (2023). The private costs of behavioral interventions. Duke Law J. 72.

Tor, A. and Klick, J. (2022). When should governments invest more in nudging? Revisiting Benartzi et al. 2017, unpublished manuscript.
Truelove, Heather B., Carrico, Amanda R., Weber, Elke U., Raimi, Kaitlin T., and Vandenbergh, Michael P. (2014). Positive and negative spillover of pro-environmental behavior: an integrative review and theoretical framework. Global Environ. Change 29: 127–138.
Vecchio, R. and Cavallo, C. (2019). Increasing healthy food choices through nudges: a systematic review. Food Qual. Prefer. 78: 1–11.
Weber, E.U., Baron, J., and Graham, L. (2001). *Conflict and tradeoffs in decision making*. Cambridge University Press, Cambridge.
Weimer, D.L. (2017). *Behavioral economics for cost-benefit analysis: benefit validity when sovereign consumers seem to make mistakes*. Cambridge University Press, Cambridge, UK.
Wilson, J.O. (1989). *Bureaucracy: what government agencies do and why they do it*. Basic Books, New York, NY.
World Health Organization (2017). *WHO report on the Global Tobacco Epidemic, 2017: monitoring tobacco use and prevention policies*. World Health Organization, Geneva.
Zamir, E. (1998). The efficiency of paternalism. Va. Law Rev. 84: 229–286.
Zarghamee, H.S., Messer, K.D., Fooks, J.R., Schulze, W.D., Shang, W., and Jubo, Y. (2017). Nudging charitable giving: three field experiments. J. Behav. Exp. Econ. 66: 137–149.