RIVISTA INTERNAZIONALE DI FILOSOFIA E PSICOLOGIA
DOI: 10.4453/rifp.2022.0017
ISSN 2039-4667; E-ISSN 2239-2629
Vol. 13 (2022), n. 3, pp. 175-186

RICERCHE

Mechanistic explanations and the ethics of nudging

Stefano Calboli[(α)] & Vincenzo Fano[(β)]

Ricevuto: 31 gennaio 2022; accettato: 27 dicembre 2022

Abstract Nudges have proven to be effective tools for steering citizens toward desirable behaviors and make valuable additions to any policy-maker's toolbox. Disappointingly, however, there are no mechanistic explanations for how nudges work, leaving policy-makers unable to explain what happens when they are implemented. This paper identifies some neglected ethical implications of the resulting citizens lack of awareness of such mechanisms. We first examine mechanistic explanations in relation to citizens' understanding on how they work. Then, we look at mechanistic explanations in light of the suggestion advanced by some ethicists that nudges be considered ethically acceptable in modern liberal democracies provided the explicit transparency of the nudges employed.
KEYWORDS: Nudge; Ethics of Nudging; Mechanistic Evidence; In-principle Transparency; Explicit Transparency

Riassunto *Spiegazioni meccanicistiche e l'etica dei* nudges – I *nudge* si sono rivelati strumenti di *policy* efficaci nello "spingere gentilmente" i cittadini verso comportamenti considerati desiderabili. Per questa ragione i *nudge* sono considerati validi strumenti della cassetta degli attrezzi di un *policymaker*. Tuttavia, è in qualche modo deludente la mancanza di spiegazioni meccanicistiche dei *nudge*, dalla quale risulta l'incapacità dei *policymaker* di spiegare quello che succede quando i *nudge* vengono effettivamente implementati. Questo articolo identifica alcune implicazioni etiche finora trascurate connesse alla inconsapevolezza da parte dei cittadini circa i meccanismi che governano i *nudge*. Da una parte, le nostre considerazioni metteranno in relazione le spiegazioni meccanicistiche con la comprensione dei *nudge* da parte dei cittadini. Dall'altra parte, metteremo in relazione le spiegazioni meccanicistiche con il concetto di trasparenza esplicita, ossia la richiesta avanzata da alcuni eticisti al fine di considerare i *nudge* eticamente accettabili nelle moderne democrazie liberali.
PAROLE CHIAVe: Nudge; Etica del nudging; Evidenza meccanicistica; Trasparenza in via di principio; Trasparenza esplicita

[(α)]Centro de éthica, polìtica e sociedade, ILCH - Universidade do Minho, Campus de Gualtar – Braga 4710-057 (PT)
[(β)]Dipartimento di Scienze Pure e Applicate, Università degli Studi "Carlo Bo", via Viti, 10 – 61029 Urbino (IT)

E-mail: calbolistefano@gmail.com (✉); vincenzo.fano@uniurb.it

NUDGES ARE POLICY TOOLS EMPLOYED worldwide. Since the US government recruited Cass Sunstein to be head of the OIRA in 2009, and the UK established the Behavioural Insights Team in 2010, nudge units have been popping up all over the world.[1] Nudge units are meant to advise on high-performing and cost-effective public policies.[2] These units take advantage of interventions that leverage human biases in choice environments. Nudges commonly refer to interventions that do not involve any form of "hard" policy: coercion, bans, or significant economic (dis)incentives. Instead, they simply ensure that citizens have the opportunity to deviate easily from the outcome desired by the choice architect.[3] This is the reason why nudges are often deemed to be "soft" policy tools, in contrast to traditional hard policy tools.

Given the scope of this paper, we will discuss just one of the many cognitive biases that nudges can exploit: *the default effect*. The default effect is a phenomenon whereby one option in a defined set of options is selected if the decision-maker does nothing. It is thus considered *the option by default*. The default effect is among the most robust biases observed by behavioral and cognitive scientists, and its impact has been confirmed in a wide range of contexts.[4] Nudges based on this effect steer decision-makers towards one of the available outcomes.

Although a burgeoning literature confirms the advantages of employing a variety of nudges, including those that exploit the default effect, explanations for their influence remain poor or are sometimes altogether lacking. This lack of knowledge by experts and policymakers results in the citizens' ignorance of the *mechanisms* that nudges' rely on; the citizens' unawareness of mechanisms is the focus of the present paper.

Here, we refer to "mechanisms" within the framework of knowledge enhancement. Mechanistic explanations are developed at several intermedial epistemic stages from the highest stage, where we have no explanation for the phenomenon considered to the allegedly lowest stage, where the physical entities and operations characterizing a phenomenon are described in physical-mathematical terms. Following Till Grüne-Yanoff,[5] we will consider the default effect to be at one of these unexplained stages, namely the difference-making stage, where an effect that corresponds to a specific pattern of events is recognized but still no understood. Discovering mechanisms means passing from this stage to a lower one which includes a model of the phenomenon, namely a causal explanation, that is both compatible with even lower stages and preserves all the relevant available evidence.[6]

From now on, we will refer to "mechanisms" or "mechanistic explanations" as elements at *this lower* stage, the stages *below* the difference-making stage.[7] We propose to refer to them as "explanatory stages".

Investigations into nudging mechanisms are scarce in general, but, as noted above, we will only refer to a specific case, nudges based on the default effect. We do so because Grüne-Yanoff has already examined the mechanisms that lead to the default effect in an influential paper. In his paper, Grüne-Yanoff emphasizes that, so far, our understanding of nudges is at the difference-making stage and not yet at an explanatory stage. Such lack of knowledge is not free from consequences, and is not optimal for policy-makers. Indeed, Grüne-Yanoff points out that if we better understood these mechanisms, we could distinguish the conditions in which nudges are likely to fail form those in which nudges succeed,[8] that is when the ascertained strength of a nudge established within a given context, e.g., under ideal experimental conditions, can be reasonably expected to replicate in different contexts. We could refer to this issue as the "effectiveness" of nudges.

Grüne-Yanoff breaks up the concept of "effectiveness", distinguishing "robustness" from "persistence". "Robustness" is diminished when a slight modification of the initial conditions will affect secondary factors, which, in turn, could mitigate, delete, or even reverse the effect on the target factor. "Persistence" is diminished if the desired consequences reduce with repetition, resulting in a significant loss of strength over the long run.

Grüne-Yanoff does not only discuss practical issues related to our poor understanding of nudging mechanisms. Turning to ethically relevant concerns, Grühe refers to the concept of "welfare", defending an approach whereby the welfare resulting from a policy depends both on the outcome and the process through which that outcome is reached.[9] In particular, he asks that we consider the degree of deliberation involved in devising and implementing a policy: typically, the higher the degree, the more desirable the policy. We believe this point deserves deeper investigation.[10] Indeed, the greater the value placed on the deliberation process, the more important it becomes for policymakers to understand the degree of manipulation a nudge involves and perform suitable procedures to ensure adequate public scrutiny of nudges. We will develop this point in §2.

While, on the one hand, Grüne-Yanoff [11] recognizes that the elements considered in mechanistic models could be described either in neurophysiological or mental terms, he mainly uses the latter. By constrast, Felsen and Reiner[12] focus on the role neuroscience could play in elucidating the mechanisms that underpin nudges. They argue that neuroscience is a valuable ally in seeking answers to empirical questions on nudging, hence in gaining insight into how nudges work and how their use could be enhanced. They further highlight how evidence from neuroscience could prove fruitful in answering open *normative* questions. The authors

claim that neuroscience could cast light on whether nudges *factually* disrespect decision-making autonomy. They consider the concept of autonomy to be defined by specific factors, such as the rationality of the choice and whether the level of autonomy is both in line with higher-order desires and free from *undue* external influences. Secondly, they argue that casting light on the neural bases of nudges will reveal when, in general, the factors that define autonomy are in place. In fact, neuroscience leads us to believe that being subject to undue external influences is «the norm rather than the exception [insomuch as such influences] are often incorporated into our decision processes covertly».[13] Hence, when considering undue external influences, it would be inaccurate to consider nudges unethical *because* they violate autonomy.

We recognize the observations advanced by Grüne-Yanoff and Felsen and Reiner as of paramount importance in both indicating the consequences of our lack of knowledge on nudging mechanisms when such mechanisms are employed, and the extent to which this ignorance matters for the ethics of nudging.[14] In line with these considerations, we believe that developing tenable models for the mechanisms underlying nudges, or, to put it in another way, replacing difference-making placeholders with mechanistic explanations would enlarge, enrich, and reframe the ethical debate on nudging at a much higher level than has so far been considered.

The gist of this paper consists in presenting further ways, so far overlooked, in which mechanistic explanations could affect the debate on the ethics of nudging. We begin by providing the necessary tools to delve into the lack of mechanistic explanations for the default effect, then present three viable candidate mechanisms for the default effect and two examples of nudges based on this effect (§1). We proceed by arguing that opening the black box of the effects exploited by nudges and disseminating the resulting knowledge would mitigate citizens' tendency to veto the use of nudges (§2). Then, we consider how knowledge of mechanistic explanations can put citizens in a position to fully control and assess government policies (§2). The fourth and fifth aspects we discuss concern the explicit transparency of nudges. In light of their allegedly subtle nature, some scholars consider transparency to be a crucial ethical condition for the use of nudges. We will argue that the knowledge obtained through access to the "explanatory stage" would impact both the evaluation of the conditions under which explicit transparency should be requested (§3) and the very feasibility of explicit transparency (§4). In what follows, we focus exclusively on the default effect, but our considerations hold for all nudges for which we lack a mechanistic explanation for their effects.

1 Nudged based on default. Three viable candidates as mechanisms

To expose the nature of the default effect, let us consider the following two examples. Firstly, a pension plan: *SMarT*. *SMarT, which* stands for *Save More Tomorrow*, is a 401(k) pension plan[15] devised by Richard Thaler and Shlomo Benartzi[16] that takes advantage of a combination of nudges in order to counteract US citizens low tendency to save. *SMarT* recognizes that participants prefer to commit themselves to start saving in the *future* rather than in the present. This contrasts with the so-called *present bias*, namely the human tendency to prefer less rewarding short-term opportunities to better long-term opportunities.[17] Acknowledging this bias, *SMart* participants are asked to increase saving only when pay raises occur. The aim is to take advantage of loss aversion, namely the human tendency to perceive the pain of losing (what is already owned) to be twice as powerful as the pleasure of gains.

While *SMarT* allows participants to opt-out of the plan as they please, the default option is to opt-in. This nudges participants to stick with the program, capitalizing on the default effect. Henceforth, we refer to this nudge, an integral component of *SMarT,* as *SMarT-by-default*. *SMarT* is among the most impressive successes obtained by behavioral economists. It has been estimated that it has aided millions of Americans to increase their savings. *SMarT* was so successful that American lawmakers built on the cognitive strategies applied in *SMarT* in designing the 2006 Pension Protection Act.[18]

Vaccination provides a second example of the default effect. The salient choice environment for vaccine appointments could be set up in two ways. On the one hand, choice architects may offer vaccine appointments as opt-in options, *viz.,* citizens are requested to *actively* make an appointment based on the available slots. Here, the default option is to not have an appointment, and thus avoid vaccination. On the other hand, choice architects could set up appointments for the citizens asking those who are not interested to actively opt out of their appointment. Here, by default, citizens end up with an appointment. It has been shown that the latter choice environment produces a higher number of vaccine appointments, and, in turn, increases the probability citizens will be vaccinated.[19] Henceforth, we will refer to this nudge as the *vaccine-appointment-by-default*.

Both these nudges are based on the default effect and are considered effective in shaping choices. But *why* do the *SMartT-by-default* and *vaccine-appointment-by-default* strategies work? How can we explain the power exerted by the default effect? Answers to such questions are useful for identifying cases in which nudges can be reasonably expected to be successful. Unfortunately, there are

no satisfying answers not only for the default effect, but more generally for all known effects on which nudges are based. Although disappointing, this should not take us by surprise; indeed, psychologists tend to consider theories as verbal constructs that organize experimental regularities.[20] Behavioral policies are based on cognitive and behavioral psychology research and rely more on behavioral data than on rigorous psychological explanations. This is reflected by the fact that random controlled trials are typically prioritized among the several evidence-generating methods available to support behavioral policies, and models to explain the phenomena observed and exploited, are typically neglected.

However, scholars are not flying completely blind when it comes to mechanisms; they can develop hypotheses that will help them move from the difference-making stage to the explanatory stage.

Considering the default effect, Grüne-Yanoff suggests that at least three cognitive mechanisms should be considered as viable candidate mechanisms for the default effect. The first candidate is "cognitive effort". The hypothesis is that the default option involves minimal cognitive effort, in contrast to the alternative options which would involve, for instance, retrieving information, a detailed comparison of final outcomes, and emotional arousal.[21] Henceforth, we will refer to this mechanism as the *cognitive-effort-mechanism*. We will refer to the second viable candidate as the *loss-aversion-mechanism*. A default determines, within a given decision context, what a decision-maker perceives the reference point to be. By setting the default, choice architects influence what decision-makers perceive to be relative losses or gains This could influence decisions because deviations of a defined magnitude from the reference point have a higher psychological impact when they are perceived as losses rather than gains.[22] The fact that losses hurt us more than gains make us happy seems to result from our evolutionary history.[23] Regarding savings: «Setting a high retirement fund contribution as the default lets the chooser interpret alternative choices with *lower contribution rates as a gain in current consumption* and a *loss in future financial security*. Consequently, according to the loss aversion model, she will put more weight on the financial security under this default than she would if the default were a low contribution».[24]

Finally, the *recommendation effect* might be the third mechanism behind the allure of default options. According to this explanation, the default effect emerges because by setting a default option, the choice architect flags it as the best option. Citizens infer that it must be in their own best interests to stick with the default.

In the following section, we will refer to the *cognitive-effort-mechanism*, *loss-aversion-mechan-ism*, and *recommendation-effect-mechanism* as we discuss how mechanistic explanations impact citizens' ability to evaluate the real-world effects of nudges.

2 The impact of mechanistic explanation on public scrutiny

In this section, we discuss how insights into the mechanisms behind nudges might shape citizens' ethical perspectives on the nature of human preferences. We argue that mechanistic explanations allow citizens to fully appreciate the potential effects of nudging. Our second point, which is deeply aligned with Felsen and Reiner's work on how nudges allegedly jeopardize human autonomy, is that mechanistic explanations can blur the perceived border between choice environments which feature nudges and nudge-free choice environments.

Being clueless on nudging mechanisms can lead citizens to infer that the effects behind nudges are specific to nudging contexts. Assuming that these kinds of mechanisms and effects are in place exclusively when nudges are implemented leads citizens to infer that nudging, in which the intentional shaping of the choice environment affects their behaviours, is somehow an exception. If they were aware of the mechanisms that support nudges, they would instead see that nudges are effective because they rely on pervasive cognitive traits related to decision-making, which do not necessarily involve the intentional interventions of external agents. If citizens were aware that the mechanisms responsible for nudging effects are pervasive, they would question their naive belief in their full freedom of choice and embrace a different perspective, recognizing the malleability of human preferences. Arguably, this would lead citizens to soften their prejudicial tendency to veto the deployment of nudges, sometimes based on the belief that nudges are characterized by unusual intrusiveness.[25]

To give an example, let us assume that loss aversion turns out to be the mechanism behind the default effect. If so, knowledgeable citizens would recognize how several of the choice environments they inhabit are, most often due to accidental circumstances, featured in a way in which loss aversion affects decisions. Loss aversion is indeed considered to be a ubiquitous tendency – within certain boundaries[26] – across a wide variety of contexts, including law-making,[27] stock investments,[28] and voting.[29] Loss aversion concerns both choices made under certainty, and under conditions of risk; loss aversion is indeed the tendency at the base of Kahneman and Tversky's successful model of risky choices.[30]

A second reason why mechanistic explanations are relevant for the ethics of nudges concerns public scrutiny of the desired behavioral outcomes for which nudges are employed in the first place. So far, we have considered the impact of mechanistic ex-

planations on the evaluation of the *process* of nudging, however, interesting ethical considerations also arise with respect to *behaviors* produced by nudges.

Moreover, mechanistic explanations would give citizens the information they need to assess the strategies and agendas used by policy-makers. On the one hand, policy-makers employ nudges to reach behavioral aims and thereby change society. From this perspective, nudges are exactly like traditional policy tools, such as bans, coercive measures, and economic incentives. On the other hand, in modern liberal democracies, public scrutiny of institutional activities is of primary concern for the good functioning of democracies and a key issue in institutional design. As part of public scrutiny activities, citizens should be able to evaluate whether policy-makers are justified in expecting a certain policy will be reasonably effective given certain circumstances.[31] We argue that casting light on the mechanisms, which are relevant for the success of nudges, could expand and improve the ability of citizens to oversee the policy-makers' work.

To see why, we could consider a hypothetical case in which the *recommendation-effect-mechanism* has been found to underpin the default effect. Let us assume that a hypothetical policy-maker has implemented a policy of *vaccine-appointment-by-default* to boost vaccine uptake. In this case, the trust placed in the policy-maker, namely the perceived source of the nudge, should be expected to play a pivotal role in the success of the nudge. Indeed, a flourishing literature on vaccination and similar health decisions shows that the degree of trust placed in the source of interventions or recommendations is pivotal in ensuring compliance. For instance, trust placed in messengers is a moderating factor in decision processes and will lead citizens to obtain health information through messages to responsibly engage in protective behaviors.[32] Furthermore, trust, or rather, a lack of it, has recently been proposed as an alternative explanation for vaccine hesitancy in contrast to the "war on science" narration which emphasizes the role of scientific illiteracy.[33]

So, trust in choice architects would be pivotal in the nudge's success if the *recommendation-effect-mechanism* was found to underpin the default effect and, in turn, *vaccine-appointment-by-default*. If so, knowledgeable citizens who scrutinize the work made by policy makers would be in the condition to consider *vaccine-appointment-by-default* as misplaced if implemented by a choice architect mistrusted by the vast majority of citizens. Instead, other things being equal, if the choice architect who takes advantage of the *vaccine-appointment-by-default* nudge was considered highly trustworthy, knowing the mechanism behind the nudge would lead citizens to expect a high degree of success. However, if loss aversion is found to be the mechanism that underpins the default effect, trust placed in the source should play a negligible role and so the source would be disregarded by citizens in their evaluation of the effectiveness of the nudge.

To summarize, a citizen unaware of the mechanisms behind nudges is unaware of the degree of success a nudge is likely to have. This, in turn, limits citizens' ability to monitor and properly assess the government's work and so engage in appropriate public scrutiny. In other words, citizens' abilities to scrutinize the work and the tenability of the assumptions made by public choice architects are made worse because nudging mechanisms remain unknown.

In this section, we have argued that mechanistic explanations are relevant in both citizens' evaluation of the moral justifiability of the processes behind the implementation of nudges and citizens' ability to perform public scrutiny. In what follows, we will focus on how an awareness of nudging mechanisms interacts with the ethical request to reveal the employment of nudges through explicit statements. This is an interesting topic in that many scholars believe that making nudges explicitly transparent is ethically necessary; they must be actually detectable by citizens and, as a result, employable in modern liberal democracies.

3 Mechanistic explanations and the need for explicit transparency

The introduction of mechanistic explanations would likely impact the debate on both the need for and the factual feasibility of explicit transparency in nudging. With "explicit transparency", we mean the addition of a statement where the presence of the nudge within the choice environment is spelt out. The main ethical concern that has led scholars to discuss and request "explicit transparency" is the allegedly subtle nature of nudges. To fully comprehend why some scholars consider nudges to be subtle, we need to consider the historical background of nudge theory. The availability of nudges as policy tools can be traced back to investigations on human "bounded rationality", pioneered by Herbert Simon.[34] Simon paved the way to systematically investigate how human cognitive processes result in predictable deviations from the "*homo oeconomicus*" model. The *homo oeconomicus* is an intentionally highly idealized model[35] for which agents are considered hyperrational, utility-maximizing, and completely self-regarding.[36] Daniel Kahneman developed this line of research and advanced a now-famous dual-system theory of mind (Kahneman et al. 1982).[37] In a nutshell, this theory describes human decisions as resulting from cognitive processes ascribable to two systems working in parallel, which should be understood as ideal types that lack precisely corresponding neuronal correlates, namely the fast "System 1" and the slow "System 2".[38] Sys-

tem 1 is a sort of unconscious pilot, which rushes us to make *automatic*, quick, cognitively effortless, and typically unaware decisions. In contrast, System 2 pushes humans to make decisions slowly, deliberatively, and with cognitive effort. Although the, so to speak, division of labor between System 1 and System 2 most often succeeds in leading us to appropriate decisions there are circumstances in which decision-makers rely solely system 1 when an intervention from System 2 was required.[39] For instance, relying on System 1 to calculate 17 x 24 would most likely lead to the wrong answer since we would fall prey to a plethora of cognitive biases.[40] Since nudges exploit cognitive biases that result from automatic and typically subpersonal cognitive processes, some scholars consider them to be subtle. This implies they are highly likely to go unnoticed by citizens. Being tendentially undetectable, nudges jeopardize citizens' decision-making autonomy.[41]

However, explicit transparency has not been the only solution advanced within the debate on the transparency of nudges. In a seminal paper, Luc Bovens[42] argued for the *in-principle transparency* of Thaler-Sunstein-style nudges. Nudges – Bovens argues – are in-principle transparent because they can be detected by people who take full advantage of their cognitive abilities. Let us consider the *vaccine-appointment-by-default* nudge to see how "in-principle transparency" is supposed to work. In this case – Bovens would likely argue – *watchful* citizens recognize that policy-makers set vaccine appointments by default to exploit the default effect in order to boost appointments. This in-principle transparency makes nudges different from subliminal messages, whose existence remains concealed, no matter what cognitive effort is made.[43] Imagine, for instance, a scenario in which evil governments force television broadcasters to occasionally flash the message "make an appointment for your vaccine". In this case, citizens watching the broadcast would be unable to detect the message, regardless of the cognitive effort they made.

Hence, in Bovens' terms, nudges are in-principle-token-transparent, so they are policy tools whose here-and-now interferences can be detected by watchful citizens.

For our part, we doubt that being watchful is sufficient to, in practice, detect nudges. We believe that citizens are in need of specific knowledge about the salient traits of nudges to factually detect them within a complex choice environment and that it is hard to argue that citizens have a duty to obtain such awareness. Being watchful seems only one condition, albeit indispensable. For instance, let us consider the complexity of the choice environment of which the *SMarT-by-default* nudge forms part. Firstly, employees must manage to distinguish between implemented nudges and the irrelevant features of the retirement plan, such as the range of fund options available. Furthermore, employees should be able to disentangle the *SMarT-by-default* nudge from the other nudges featured in *SMarT*, for instance the nudge meant to contrast the present bias. We suggest that employees would only be able to factually distinguish the *SMarT-by-default* nudge within the choice environment if they had previous educated themselves about the default effect.

Our claim is supported by literature on the crucial role played by education (long-term process) and training (short-term process) in recognizing the threats of biased cognitive processes and shielding ourselves from them. The cases in which lurking cognitive biases can be recognized and defused through education and training are *analogous* to those in which education could influence the probability that nudges could be detected and eventually withstood.[44]

For this and other reasons, many scholars have begun to empirically investigate the feasibility of making nudges *explicitly* transparent.[45] As things stand, the debate seems to be polarized into two competitive positions. On one side are those who assume that the in-principle transparency of nudges suffices to comply with the ethical demands of modern liberal democracies: on the other side, those who argue that nudges must be made explicitly transparent before they can be considered suitable policy tools.

We argue that opening the black box to examine the effects that underpin nudges and casting light on the mechanisms underlying these effects would result in more flexible positions within the debate on transparency. Mechanistic explanations would lead scholars to recognize the need to assess the kind of transparency needed on a case by case basis. Let us develop this argument considering the *vaccine-appointment-by-default* nudge. On the one hand, let us depict a hypothetical scenario in which we collected evidence indicating the *recommendation effect* to be the mechanism underlying the default effect. This would imply that citizens, when successfully nudged, would necessarily be able to recognize the presence of the *vaccine-appointment-by-default* nudge and the behavior this nudge is designed to steer, so that they could deliberately either accept or reject the suggestion. Such a piece of evidence would arguably lead scholars who root for explicit transparency to recognize that there are indeed cases in which asking for explicit transparency would be pointless, because the mechanism underlying the nudge shows that the nudge is inherently detectable when effective.

On the other hand, let us assume that *loss-aversion-mechanism* underpins the *vaccine-appointment-by-default* nudge. If so, citizens would face a nudge, whose mechanism is based on an automatic response which is difficult to detect. Such

considerations would arguably lead scholars who generally claim it is unnecessary to ask for explicit transparency to make an exception, recognizing the need for explicit transparency for nudges characterized by high undetectability. Mechanistic explanations would thus enable us to move from a dispute characterized by two all-encompassing positions to a debate in which scholars, albeit maintaining different perspectives, recognize a nuanced situation in which a case by case analysis is necessary.

In this section we have discussed how mechanistic explanations affect perspectives on explicit transparency as an ethical condition of nudging. In the next section, we will take a step back and discuss whether making citizens aware of nudging mechanisms, together with explicit transparency, would undermine the strength of nudges, to the extent that nudging might even become pointless.

4 Mechanistic explanatins and the feasibility of explicit transparency

We now consider the possibility that awareness of nudging mechanisms could also impact the *feasibility* of making nudges explicitly transparent. We consider explicitly transparent nudges feasible when the information meant to make nudges detectable does not heavily impair their strength in shaping citizens' behaviors. In this regard, Luc Bovens claimed that nudges typically work better in the dark.[46] Furthermore, as mentioned by Loewenstein and colleagues,[47] in the Behavior Change report, the UK House of Lords discussed explicit transparency in nudging, concluding that while ensuring transparency, is ethically preferable, it is not the most suitable solution since «this fuller sort of transparency might limit the effectiveness of the intervention».[48]

The main concern among policy-makers and scholars regarding explicit transparency in nudging is psychological reactance. Psychological reactance is a concept introduced first by Brehm,[49] and it consists in the «unpleasant motivational arousal that emerges when people experience a threat to or loss of their free behaviors [... this...] results in behavioral and cognitive efforts to reestablish one's freedom».[50]

If nudges are in fact perceived as a threat to freedom of choice,[51] they could trigger reactance and grow weak, or, in the worst-case scenario, even backfire. This eventuality needs to be empirically investigated, and scholars have recently begun to do so.

It appears that our seemingly reasonable intuition that transparency could compromise the power of nudges and trigger psychological reactance stands on empirically shaky ground. The evidence on the consequences of making nudges explicitly transparent are mixed[52] and, although this does not clearly indicate that explicitly transparent nudges are totally feasible, it surely does not suggest the opposite.

In the case of the default effect, it seems that the explicit transparency of nudges does not significantly impair their strength. To the best of our knowledge, the research conducted so far on default-based nudges indicates that explicit transparency is in fact feasible.[53] Taken together, these investigations consider three strategies for making nudges transparent and several combination approaches. They examine the effect of transparency on: the *behavioral result* of the nudge; the *effect exploited*; and, eventual *side-effects* due to the nudge. Let us see in detail what these three strategies entail. With respect to the *vaccine-appointment-by-default*, transparency regarding the behavioral result would consist in a statement of this kind: "Previous research shows that providing appointments by default results in a greater number of appointments than if having an appointment is not the default option". Instead, the transparency of the effect exploited could be guaranteed by a statement like: "The default effect makes the option-by default more attractive to decision-makers". Finally, making possible side-effects transparent would mean warning that the nudge will discourage some citizens, in this case, those who refuse vaccines, from obtaining their favored outcome. In all three strategies, the nudges' strength seems to remain intact. Let us now consider the eventual explicit transparency of the mechanism.

We argue that if knowledge of the nudging mechanism is at citizens' disposal, it would intertwine with those aspects of nudges that have been made explicitly transparent and, in turn, impact on their effectiveness. To see why, let us consider two slightly different scenarios. In both of them, citizens inhabit the very same choice environment where the *vaccine-appointment-by-default* nudge is in place. Furthermore, in both scenarios, the nudge is made explicitly transparent.

For the sake of argument, it is irrelevant what aspect of the nudge is made more salient through transparency: the behavioral result, the effect exploited, or/and the potential side-effect. What matters here is that citizens are, through explicit transparency, made aware of the presence of a nudge. Let us assume that, in the first scenario, knowledge of the mechanism behind the effect on which the *vaccine-appointment-by-default* nudge is based is not yet available. Conversely, in the second scenario citizens have access to the knowledge of the mechanism behind the *vaccine-appointment-by-default* nudge.

We argue that in the second scenario, citizens awareness of nudging mechanisms would impact the perception of the nudges themselves, eventually allowing psychological reactance to be triggered. This, in turn, could undermine the feasibility of explicitly transparent nudges. Indeed, the knowledge of the mechanism characterizing the second scenar-

io would arguably make the degree of manipulation entailed by the nudge more salient. By assumption, in the first scenario, this aspect is absent.

Among the factors that intensify psychological reactance is a perceived intention to persuade which could be triggered by dramatic narratives[54] or the use of forceful and controlling wording.[55] More specifically, with regard to nudges, it seems likely that being aware of the mechanisms behind the nudges could be a further factor in triggering the perception of a strong intention to persuade, depending on the mechanistic explanation in place.

Let us depict a case in which the mechanism behind the *SMarT-by-default* nudge turns out to be the *loss-aversion-mechanism*. In this case, the detected nudge would be perceived as a policy tool implemented by the choice architect due to its ability, through prompting of automatic responses with an early-evolutionary origin, to strongly persuade choosers. Arguably, this could intensify psychological reactance, perhaps to the extent of making the *SMarT-by-default* nudge unavoidably unsuccessful. On the other hand, it is reasonable to expect that this would not happen if the mechanism behind the nudge turned out to be the *recommendation effect*, which instead would lead citizens to perceive the nudge as a suggestion and, in turn, not a tool of persuasion. To conclude, we argue that when citizens have access to the knowledge of nudging mechanisms, this – depending on which mechanism is, in fact, revealed – could lead to psychological reactance. However, this is merely a conjecture and experimental investigations would be necessary to test if our hypothesis is tenable. The work on the perception of nudges by Michaelsen[56] could inspire such experimental investigations.

5 Conclusion

In the foregoing sections, we analyzed how and why mechanistic explanations are relevant for the ethics of nudging. To summarize, we argued that being aware of the mechanisms behind nudges could affect citizens' perception of the nature of the processes involved and enhance the public's capacity for scrutiny. Furthermore, we argued how upgrading from a difference-making stage to an explanatory stage is relevant in assessing the need for explicit transparency in nudging and its feasibility. We conclude this paper by presenting two further considerations that we hope will stimulate future research on mechanistic explanations and their impact on the ethics of nudging.

The first concerns how such mechanisms can be discovered. In previous sections, we considered three viable candidate mechanisms behind the default effect. Furthermore, we assumed that the actual mechanism is a single one, specifically, either the *cognitive-effort-mechanism*, the *loss-aversion-mechanism*, or the *recommendation-effect-mechanism*. However, this assumption could turn out to be inaccurate. Indeed, we cannot rule out the possibility that upgrading from the "difference-making stage" to the "explanatory stage" will make it evident that the default effect is, in fact, more heterogeneous than previously thought. It could be the case that various nudges based on the default effect will turn out to depend on different kinds of mechanisms or even a combination of them. This concrete possibility should further impel scholars and policy-makers who care about the success of nudges to cast light on their mechanisms.

Secondly, we discussed how an awareness of nudging mechanisms and the feasibility of explicit-transparent nudges are intertwined (§4). We left aside the case in which the mechanism itself is considered to be explicitly transparent. This is the case when, alongside the transparency of the behavioral result, policy-makers ensure other aspects are also transparent: the effect exploited, the side-effects involved, and/or the nudging mechanism. The impact of this kind of transparency on the feasibility of explicit transparency may not be trivial. This is a relevant topic in light of what we argued in §2 about public scrutiny. Indeed, if mechanistic explanations play a pivotal role in enhancing citizens' capacity to perform public scrutiny, a fundamental function in our democracies, it is reasonable to claim that the transparency of nudging mechanisms is an advisable addition from an ethical standpoint. Nevertheless, this leaves open the question of whether such transparency could heavily undermine the impact of nudges. Once more, we are dealing with a question that scholars can address only through empirical investigation.

The explanation-neglecting approach that currently characterizes the development of behavioral policies can be changed. We hope that, in the near future, cognitive scientists will become increasing aware that mechanistic explanations of mental phenomena are ethically important and will fully recognize the downsides of disregarding such explanations. Cognitive scientists should investigate mechanistic explanations for nudging and cast light on the ethical implications of such explanationsm.

Notes

[1] Cf. OECD, *Behavioural insights and public policy*.

[2] Cf. S. BENARTZI, J. BESHEARS, K.L. MILKMAN, C.R. SUNSTEIN, R.H. THALER, M. SHANKAR, W. TUCKER-RAY, W.J. CONGDON, S. GALING, *Should governments invest more in nudging?*.

[3] Cf. R. H. THALER, C. SUNSTEIN, *Nudge: The final edition*.

[4] Cf. J.M. JACHIMOWICZ, S. DUNCAN, E.U. WEBER, E.J. JOHNSON, *When and why defaults influence decisions: A meta-analysis of default effects*.

[5] Cf. T. GRÜNE-YANOFF, *Why behavioural policy needs mechanistic evidence*.

[6] Cf. M.J. NATHAN, *Black boxes: How science turns ignorance into knowledge*; G. FELSEN, P.B. REINER, *What can neuroscience contribute to the debate over nudging?*. Our

perspective on mechanistic explanations should be large enough to be compatible with both realism and anti-realism approaches to models and, as well, compatible with the several definitions of causality present in the philosophical literature.

[7] Grüne-Yanoff calls "mechanistic evidence" what we refer to as "mechanistic explanation". We consider "mechanistic explanation" more suitable in that this locution emphasizes the explanatory character of mechanisms.

[8] Cf. M. OSMAN, S. MCLACHLAN, N. FENTON, M. NEIL, R. LÖFSTEDT, B. MEDER, *Learning from behavioural changes that fail*.

[9] For instance, republicans assess negatively those processes that imply arbitrary power exerted by policymakers, being deemed as limitations of citizens' freedom, unacceptable if not democratically discussed (cf. P. PETTIT, *On the people's terms: A republican theory and model of democracy*).

[10] Cf. A. BARTON, T. GRÜNE-YANOFF, *From libertarian paternalism to nudging - and beyond*.

[11] Cf. *supra*, n. 5.

[12] Cf. *supra*, n. 6.

[13] G. FELSEN, P.B. REINER, *What can neuroscience contribute to the debate over nudging?*, p. 476.

[14] As regards practical concerns due to the lack of mechanistic explanations of nudges cf. also C. MARCHIONNI, S. REIJULA *What is mechanistic evidence, and why do we need it for evidence-based policy?*. Although the authors sympathize with Grüne-Yanoff's remarks, they advance a different and promising account of mechanistic explanations defined as "componential difference-makings". Furthermore, Grüne-Yanoff and colleagues consider mechanisms as means to distinguish between nudges and boosts (cf. T. GRÜNE-YANOFF, C. MARCHIONNI, M.A. FEUFEL, *Toward a framework for selecting behavioural policies: How to choose between boosts and nudges*).

[15] US retirement plans can be of two kinds. The first includes *Defined Benefit Pension* plans, whereby the retirement is predetermined and depends on a pension benefit formula. The second comprises *Defined Contribution* plans whereby a trust fund is created, and the retirement depends on the amount invested in the fund by the employees. 401(k), named after a subsection of the Internal Revenue Code, is a Defined Contribution plan.

[16] Cf. R.H. THALER, S. BENARTZI, *Save more tomorrow: Using behavioral economics to increase employee saving*.

[17] Cf. T. O'DONOGHUE, M. RABIN, *Present bias: Lessons learned and to be learned*.

[18] Cf. R.H. THALER, S. BENARTZI, *Behavioral economics and the retirement savings crisis*.

[19] Cf. G.B. CHAPMAN, M. LI, H. COLBY, H. YOON, *Opting in vs opting out of influenza vaccination*.; B.A. LEHMANN, G.B. CHAPMAN, F.M. FRANSSEN, G. KOK, R.A. RUITER, *Changing the default to promote influenza vaccination among health care workers*.

[20] Cf. C. CAMERER, *Behavioral economics: Reunifying psychology and economics*.

[21] Cf. E.J. JOHNSON, D. GOLDSTEIN, *Do defaults save lives?*.

[22] Cf. A. TVERSKY, D. KAHNEMAN, *Loss aversion in riskless choice: A reference-dependent model*.

[23] Cf. M. CHEN, V. LAKSHMINARAYANAN, L. SANTOS, *How basic are behavioral biases? Evidence from capuchin monkey trading behavior*.

[24] T. GRÜNE-YANOFF, *Why behavioural policy needs mechanistic evidence*, p. 469 - italics added.

[25] Cf. G. FELSEN, N. CASTELO, P.B. REINER, *Decisional enhancement and autonomy: Public attitudes towards overt and covert nudges*; C.R. SUNSTEIN, *Do people like nudges?*.

[26] Cf. N. NOVEMSKY, D. KAHNEMAN, *The boundaries of loss aversion*.

[27] Cf. J.J. RACHLINSKI, A.J. WISTRICH, *Gains, losses, and judges: Framing and the judiciary*.

[28] Cf. S. BENARTZI, R.H. THALER, *Myopic loss aversion and the equity premium puzzle*.

[29] Cf. A. ALESINA, F. PASSARELLI, *Loss aversion in politics*.

[30] Cf. D. KAHNEMAN, A. TVERSKY, *Prospect theory: An analysis of decision under risk*.

[31] We refer to citizens' evaluations *ex-ante*; hence evaluations made before policies have revealed their factual degree of success. These evaluations are likely driven by rational reflections, instrumental for the good functioning of our democracies. Instead, *ex post facto* evaluations tend to fall victim to the hindsight bias, so they cannot be considered parts of worthy public scrutiny (cf. B. FISCHHOFF, *Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty*).

[32] Cf. M.J. FIGUEIRAS, J. GHORAYEB, M.V.C. COUTINHO, J. MARÔCO, J. THOMAS, *Levels of trust in information sources as a predictor of protective bealth behaviors during COVID-19 pandemic: A UAE cross-sectional study*; R.A. BLAIR, B. MORSE, L. TSAI, *Public health and public trust: Evidence from the ebola virus disease epidemic in Liberia*; O. BARGAIN, U. AMINJONOV, *Trust and compliance to public health policies in times of Covid-19*; C. BICCHIERI, E. FATAS, A. ALDAMA, A., CASAS, I. DESHPANDE, M. LAURO, C. PARILLI, M. SPOHN, P. PEREIRA, R. WEN, *In science we (should) trust: Expectations and compliance across nine countries during the COVID-19 pandemic*.

[33] Cf. M.J. GOLDENBERG, *Vaccine hesitancy: Public trust, expertise, and the war on science*.

[34] Cf. H.A. SIMON, *A behavioral model of rational choice*.

[35] Cf. D.K. LEVINE, *Is behavioral economics doomed? The ordinary versus the extraordinary*.

[36] Cf. R. FESTA, G. CEVOLANI, *Giochi di società: teoria dei giochi e metodo delle scienze sociali*; D. HAUSMAN, *Preference, value, choice, and welfare*.

[37] D. KAHNEMAN, P. SLOVIC, A. TVERSKY, *Judgment under uncertainty*. It is worth noting that the theory of mind proposed by Kahneman is only one among the dual-system theories of mind that have been advanced. For an influential alternative dual-system theory, cf. P. WASON, J. EVANS, *Dual processes in reasoning?*.

[38] An epistemological *caveat*: Kahneman himself warns against considering System 1 and System 2 as actual systems; instead, they should be considered fictitious characters that help the reader to comprehend human cognition. Cf. D. KAHNEMAN, *Thinking, fast and slow*.

[39] For a different perspective cf. G. GIGERENZER, *Why heuristics work*; G. GIGERENZER, *On the supposed evidence for libertarian paternalism*.

[40] Cf. J. BARON, *Thinking and deciding*.

[41] Cf. R. REBONATO, *Taking liberties: A critical examination of libertarian paternalism*; P.G. HANSEN, A.M. JESPERSEN, *Nudge and the manipulation of choice*; T. GRÜNE-YANOFF, *Old wine in new casks: Libertarian paternalism still violates liberal principles*.

[42] Cf. L. BOVENS, *The ethics of nudge*.

[43] It is worth noting that whether subliminal messages factually alter our decisions is a matter of debate (cf., for instance, C. TRAPPEY, *A meta-analysis of consumer choice and subliminal advertising*; E. STRAHAN, M.

ZANNA, *Subliminal priming and persuasion: Striking while the iron is hot*); however, whether or not subliminal messages are actually effective is irrelevant from our normative perspective.

[44] As regards education, cf. G.T. FONG, D.H. KRANTZ, R.E. NISBETT, *The effects of statistical training on thinking about everyday problems*. Concerning training, cf. C. MOREWEDGE, H. YOON, I. SCOPELLITI, C. SYMBORSKI, J. KORRIS, K. KARIM, *Debiasing decisions: Improved decision making with a single training intervention*; A. SELLIER, I. SCOPELLITI, C. MOREWEDGE, *Debiasing training improves decision making in the field*. In: «Psychological Science». For overviews of debiasing strategies and discussions on the limits of education and training, e.g. on the portability of knowledge in different domains, cf. J.B. SOLL, K.L. MILKMAN, J.W. PAYNE, *A user's guide to debiasing*; D. KAHNEMAN, O. SIBONY, C.R. SUNSTEIN, *Noise: A Flaw in human judgment*; K.L. MILKMAN, D. CHUGH, M.H. BAZERMAN, *How can decision making be improved?*.

[45] Cf. for instance, J. WACHNER, M. ADRIAANSE, D. DE RIDDER, *The influence of nudge transparency on the experience of autonomy*; P. MICHAELSEN, L. NYSTRÖM, T.J. LUKE, L. JOHANSSON, M. HEDESSTRÖM, *Are default nudges deemed fairer when they are more transparent? People's judgments depend on the circumstances of the evaluation*. We will discuss this literature in §4.

[46] Cf. L. BOVENS, *The ethics of nudge*.

[47] Cf. G. LOEWENSTEIN, C. BRYCE, D. HAGMANN, S. RAJPAL, *Warning: You are about to be nudged*.

[48] HOUSE OF LORDS, SCIENCE AND TECHNOLOGY SELECT COMMITTEE, *Behaviour change* (Second report 2011).

[49] Cf. J.W. BREHM, *A theory of psychological reactance*.

[50] C. STEINDL, E. JONAS, S. SITTENTHALER, E. TRAUT-MATTAUSCH, J. GREENBERG, *Understanding psychological reactance*, p. 205. For a comprehensive discussion cf. also A.M. MIRON, J.M. BREHM, *Reactance theory – 40 years later*.

[51] On this eventuality, cf. G. FELSEN, N. CASTELO, P.B. REINER, *Decisional enhancement and autonomy: Public attitudes towards overt and covert nudges*; C.R. SUNSTEIN, *Do people like nudges?*.

[52] Cf. E. KANTOROWICZ-REZNICHENKO, J. KANTOROWICZ, *To follow or not to follow the herd? Transparency and social norm nudges*; S. CASAL, F. GUALA, L. MITTONE, *On the transparency of nudges: An experiment.*

[53] Cf. H. BRUNS, E. KANTOROWICZ-REZNICHENKO, M.L. JONSSON, B. RAHALI, *Can nudges be transparent and yet effective?*; E. KANTOROWICZ-REZNICHENKO, J. KANTOROWICZ, *To follow or not to follow the herd? Transparency and social norm nudges*; S. CASAL, F. GUALA, L. MITTONE, *On the transparency of nudges: An experiment.*

[54] Cf. E. MOYER-GUSÉ, R.L. NABI, *Explaining the effects of narrative in an entertainment television program: Overcoming resistance to persuasion*.

[55] Cf. C.H. MILLER, L.T. LANE, L.M. DEATRICK, A.M. YOUNG, K.A. POTTS, *Psychological Reactance and Promotional Health Messages: The effects of controlling language, lexical concreteness, and the restoration of freedom*; B.L. QUICK, M.T. STEPHENSON, *Examining the role of trait reactance and sensation seeking on perceived threat, state reactance, and reactance restoration*. Interestingly, among the moderators of psychological reactance, there are aspects likewise salient for nudging, such as in-group/out-group dynamics (cf. V. GRAUPMANN, E. JONAS, E. MEIER, S. HAWELKA, M. AICHHORN, *Reactance, the self, and its group: When threats to freedom come from the ingroup versus the outgroup*) and the trust placed in the source (cf. H. SONG, K.A. MCCOMAS, K.L. SCHULER, *Source effects on psychological reactance to regulatory policies: The role of trust and similarity*).

[56] Cf., for instance, P. MICHAELSEN, L. NYSTRÖM, T.J. LUKE, L. JOHANSSON, M. HEDESSTRÖM, *Are default nudges deemed fairer when they are more transparent? People's judgments depend on the circumstances of the evaluation*.

Literature

ALESINA, A., PASSARELLI, F. (209). *Loss aversion in politics*. In: «American Journal of Political Science», vol. LXIII, n. 4, pp. 936-947.

BARGAIN, O., AMINJONOV, U. (2020). *Trust and compliance to public health policies in times of Covid-19*. In: «Journal of Public Economy», vol. CXCII - doi: 10.1016/j.jpubeco.2020.104316.

BARON, J. (2007). *Thinking and deciding*, Cambridge University Press, Cambridge, 4th edition.

BARTON, A., GRÜNE-YANOFF, T. (2015). *From libertarian paternalism to nudging - and beyond*. In: «Review of Philosophy and Psychology», vol. VI, n. 3, pp. 341-359.

BENARTZI, S., BESHEARS, J., MILKMAN, K.L., SUNSTEIN, C.R., THALER, R.H., SHANKAR, M., TUCKER-RAY, M., CONGDON, W.J., GALING, S. (2017). *Should governments invest more in nudging?*. In: «Psychological Science», vol. XXVIII, n. 8, pp. 1041-1055.

BENARTZI, S., THALER, R.H. (1995). *Myopic loss aversion and the equity premium puzzle*. In: «The Quarterly Journal of Economics», vol. CX, n. 1, pp. 73-92.

BICCHIERI, C., FATAS, E., ALDAMA, A., CASAS, A., DESHPANDE, I., LAURO, M., PARILLI, C., SPOHN, M., PEREIRA, P., WEN, R. (2021). *In science we (should) trust: Expectations and compliance across nine countries during the COVID-19 pandemic*. In: «PLoS ONE», vol. XVI, n. 6, Art. Nr. e0252892 – doi: 10.1371/journal.pone.0252892.

BLAIR, R.A., MORSE, B., TSAI, L.L. (2017). *Public health and public trust: Evidence from the ebola virus disease epidemic in Liberia*. In: «Social Science and Medicine», vol. CLXXII, pp. 89-97.

BOVENS, L. (2009). *The ethics of nudge*. In: T. GRÜNE-YANOFF, S.O. HANSSON (eds.), *Preference change: Approaches from philosophy, economics and psychology*, Springer, Berlin/New York, pp. 207-219.

BREHM, J.W. (1966). *A theory of psychological reactance*, Academic Press, Boston/London/New York.

BRUNS, H., KANTOROWICZ-REZNICHENKO, E., JONSSON, M.L., RAHALI, B. (2018). *Can nudges be transparent and yet effective?*. In: «Journal of Economic Psychology», vol. LXV, Issue C, pp. 41-59.

CAMERER, C. (1999). *Behavioral economics: Reunifying psychology and economics*. In: «Proceedings of the National Academy of Sciences», vol. XCVI, n. 19, pp. 10575-10577.

CASAL, S., GUALA, F., MITTONE, L. (2019). *On the transparency of nudges: An experiment*, CEEL Working Papers n. 1902.

CHAPMAN, G.B., LI, M., COLBY, H., YOON, H. (2010). *Opting in vs opting out of influenza vaccination*. In: «Journal of the American Medical Association», vol. CCCIV, n. 1, pp. 43-44.

CHEN, M., LAKSHMINARAYANAN, V., SANTOS, L. (2006).

How basic are behavioral biases? Evidence from capuchin monkey trading behavior. In: «Journal of Political Economy», vol. CXIV, n. 3, pp. 517-537.

FELSEN, G., CASTELO, N., REINER, P.B. (2013). *Decisional enhancement and autonomy: Public attitudes towards overt and covert nudges*. In: «Judgment and Decision Making», vol. VIII, n. 2, pp. 202-213.

FELSEN, G., REINER, P.B. (2015). *What can neuroscience contribute to the debate over nudging?*. In: «Review of Philosophy and Psychology», vol. VI, n. 3, pp. 469-479.

FESTA, R., CEVOLANI, G. (2013). *Giochi di società: teoria dei giochi e metodo delle scienze sociali*, Mimesis, Milano.

FIGUEIRAS, M.J., GHORAYEB, J., COUTINHO, M.V.C., MARÔCO, J., THOMAS, J. (2021). *Levels of trust in information sources as a predictor of protective health behaviors during COVID-19 pandemic: A UAE cross-sectional study*. In: «Frontiers in Psychology», vol. XII, Art.Nr. 633550 – doi: 10.3389/fpsyg.2021.633550.

FISCHHOFF, B. (1975). *Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty*. In: «Journal of Experimental Psychology: Human Perception and Performance», vol. I, n. 3, pp. 288-299.

FONG, G.T., KRANTZ, D.H., NISBETT, R.E. (1986). *The effects of statistical training on thinking about everyday problems*. In: «Cognitive Psychology», vol. XVIII, n. 3, pp. 253-292.

GIGERENZER, G. (1998). *Why heuristics work*. In: «Perspectives on Psychological Science», vol. III, n. 1, pp 20-29.

GIGERENZER, G. (2015). *On the supposed evidence for libertarian paternalism*. In: «Review of Philosophy and Psychology», vol. VI, n. 3, pp. 361-383.

GOLDENBERG, M.J. (2021). *Vaccine hesitancy: Public trust, expertise, and the war on science (science, values, and the public)*, University of Pittsburgh Press, Pittsburgh.

GRAUPMANN, V., JONAS, E., MEIER, E., HAWELKA, S., AICHHORN, M. (2011). *Reactance, the self, and its group: When threats to freedom come from the ingroup versus the outgroup*. In: «European Journal of Social Psychology», vol. XLII, n. 2, pp. 164-173.

GRÜNE-YANOFF, T. (2012). *Old wine in new casks: Libertarian paternalism still violates liberal principles*. In: «Social Choice and Welfare», vol. XXXVIII, n. 4, pp. 635-645.

GRÜNE-YANOFF, T. (2016). *Why behavioural policy needs mechanistic evidence*. In: «Economics and Philosophy», vol. XXXII, n. 3, pp. 463-483.

GRÜNE-YANOFF, T., MARCHIONNI, C., FEUFEL, M.A. (2018). *Toward a framework for selecting behavioural policies: How to choose between boosts and nudges*. In: «Economics and Philosophy», vol. XXXIV, n. 2, pp. 243-266.

HANSEN, P.G., JESPERSEN, A.M. (2013). *Nudge and the manipulation of choice*. In: «European Journal of Risk Regulation», vol. IV, n. 1, pp. 3-28.

HAUSMAN, D. (2012). *Preference, value, choice, and welfare*, Cambridge University Press, Cambridge.

HOUSE OF LORDS, SCIENCE AND TECHNOLOGY SELECT COMMITTEE, *Behaviour change* (Second report), 2011, London, United Kingdom – available at URL: https://publications.parliament.uk/pa/ld201012/ldselect/ldsctech/179/17902.htm.

JACHIMOWICZ, J.M., DUNCAN, S., WEBER, E.U., JOHNSON, E.J. (2019). *When and why defaults influence decisions: A meta-analysis of default effects*. In: «Behavioural Public Policy», vol. III, n. 2, pp. 159-186.

JOHNSON, E.L., GOLDSTEIN, D. (2003). *Do defaults save lives?*. In: «Science», vol. CCCII, n. 5649, pp. 1338-1339.

KAHNEMAN, D. (2011). *Thinking, fast and slow*, Macmillan, New York.

KAHNEMAN, D., SIBONY, O., SUNSTEIN, C.R. (2021). *Noise: A flaw in human judgment*, William Collins, London.

KAHNEMAN, D., SLOVIC, P., TVERSKY, A. (1982), *Judgment under uncertainty: Heuristics and biases*, Cambridge University Press.

KAHNEMAN, D., TVERSKY, A. (1979). *Prospect theory: An analysis of decision under risk*. In: «Econometrica», vol. XLVII, n. 2, pp. 263-292.

KANTOROWICZ-REZNICHENKO, E., KANTOROWICZ, J. (2021). *To follow or not to follow the herd? Transparency and social norm nudges*. In: «Kyklos», vol. LXXIV, n. 3, pp. 362-377.

LEHMANN, B.A., CHAPMAN, G.B., FRANSSEN, F.M., KOK, G., RUITER, R.A. (2016). *Changing the default to promote influenza vaccination among health care workers*. In: «Vaccine», vol. XXXIV, n. 11, pp. 1389-1392.

LEVINE, D.K. (2012). *Is behavioral economics doomed? The ordinary versus the extraordinary*, Open Book, Cambridge.

LOEWENSTEIN, G., BRYCE, C., HAGMANN, D., RAJPAL, S. (2015). *Warning: You are about to be nudged*. In: «Behavioral Science and Policy», vol. I, n. 1, pp. 35-42.

MARCHIONNI, C., REIJULA, S. (2019). *What is mechanistic evidence, and why do we need it for evidence-based policy?*. In: «Studies in History and Philosophy of Science – Part A», vol. LXXIII, pp. 54-63.

MICHAELSEN, P., NYSTRÖM, L., LUKE, T.J., JOHANSSON, L., HEDESSTRÖM, M. (2020). *Are default nudges deemed fairer when they are more transparent? People's judgments depend on the circumstances of the evaluation*. In: «PsyArxiv Preprints» - doi: 10.31234/osf.io/5knx4

MILKMAN, K.L., CHUGH, D., BAZERMAN, M.H. (2009). *How can decision making be improved?*. In: «Perspectives on Psychological Science», vol. IV, n. 4, pp. 379-383.

MILLER, C.H., LANE, L.T., DEATRICK, L.M., YOUNG, A.M., POTTS, K.A. (2007). *Psychological Reactance and Promotional Health Messages: The effects of controlling language, lexical concreteness, and the restoration of freedom*. In: «Human Communication Research», vol. XXXIII, n. 2, pp. 219-240.

MIRON, A.M., BREHM, J.M. (2006). *Reactance theory – 40 years later*. In: «Zeitschrift für Sozialpsychologie», vol. XXXVII, n. 1, pp. 9-18.

MOREWEDGE, C., YOON, H., SCOPELLITI, I., SYMBORSKI, C., KORRIS, J., KARIM, K. (2015). *Debiasing decisions: Improved decision making with a single training intervention*. In: «Policy Insights from the Behavioral and Brain Sciences», vol. II, n. 1, pp. 129-140.

MOYER-GUSE, E., NABI, R.L. (2010). *Explaining the effects of narrative in an entertainment television program: Overcoming resistance to persuasion*. In: «Human Communication Research», vol. XXXVI, n. 1, pp. 26-52.

NATHAN, M.J. (2021). *Black boxes: How science turns ignorance into knowledge*, Oxford University Press,

Oxford.

NOVEMSKY, N., KAHNEMAN, D. (2005). *The boundaries of loss aversion*. In: «Journal of Marketing Research», vol. XLII, n. 2, pp. 119-128.

O'DONOGHUE, T., RABIN, M. (2015). *Present bias: Lessons learned and to be learned*. In: «American Economic Review», vol. CV, n. 5, pp. 273-279.

OECD (2017). *Behavioural insights and public policy: Lessons from around the world*, OECD Publishing, Paris.

OSMAN, M., MCLACHLAN, S., FENTON, N., NEIL, M., LÖFSTEDT, R., MEDER, B. (2020). *Learning from behavioural changes that fail*. In: «Trends in Cognitive Sciences», vol. XXIV, n. 12, pp. 969-980.

PETTIT, P. (2012). *On the people's terms: A republican theory and model of democracy*, Cambridge University Press, Cambridge.

QUICK, B.L., STEPHENSON, M.T. (2008). *Examining the role of trait reactance and sensation seeking on perceived threat, state reactance, and reactance restoration*. In: «Human Communication Research», vol. XXXIV, n. 3, pp. 448-476.

RACHLINSKI, J.J., WISTRICH, A.J. (2018). *Gains, losses, and judges: Framing and the judiciary*. In: «Notre Dame Law Review», vol. XCIV, n. 2, pp. 521-582.

REBONATO, R. (2012). *Taking liberties: A critical examination of libertarian paternalism*, Palgrave-Macmillan, London/New York.

SELLIER, A., SCOPELLITI, I., MOREWEDGE, C. (2019). *Debiasing training improves decision making in the field*. In: «Psychological Science», vol. XXX, n. 9, pp. 1371-1379.

SIMON, H.A. (1955). *A behavioral model of rational choice*. In: «Quarterly Journal of Economics», vol. LXIX, n. 1, pp. 99-118.

SOLL, J.B., MILKMAN, K.L., PAYNE, J.W. (2016). *A user's guide to debiasing*, In: G. KEREN, G. WU (eds.), *The Wiley Blackwell handbook of judgment and decision making*, Wiley-Blackwell, New York/London, pp. 924-951.

SONG, H., MCCOMAS, K.A., SCHULER, K.L. (2018). *Source effects on psychological reactance to regulatory policies: The role of trust and similarity*. In: «Science Communication», vol. XL, n. 5, pp. 591-620.

STEINDL, C., JONAS, E., SITTENTHALER, S., TRAUT-MATTAUSCH, E., GREENBERG, J. (2015). *Understanding psychological reactance*. In: «Zeitschrift für Psychologie», vol. CCXXIII, n. 4, pp. 205-214.

STRAHAN, E., ZANNA, M. (2002). *Subliminal priming and persuasion: Striking while the iron is hot*. In: «Journal of Experimental Social Psychology», vol. XXXVIII, n. 6, pp. 556-568.

SUNSTEIN, C.R. (2016). *Do people like nudges?*. In: «Administrative Law Review», vol. LXVIII, n. 2, pp. 177-232.

THALER, R.H., BENARTZI, S. (2004). *Save more tomorrow™: Using behavioral economics to increase employee saving*. In: «Journal of Political Economy», vol. CXII, n. S1, pp. 164-187.

THALER, R.H., BENARTZI, S. (2013). *Behavioral economics and the retirement savings crisis*. In: «Chicago Booth Review», available at URL: http://review.chicagobooth.edu/magazine/summer-2013/retirement-savings.

THALER, R.H., SUNSTEIN, C. (2021). *Nudge: The final edition*, Penguin Books, New York.

TRAPPEY, C. (1996). *A meta-analysis of consumer choice and subliminal advertising*. In: «Psychology and Marketing», vol. XIII, n. 5, pp. 517-530.

TVERSKY, A., KAHNEMAN, D. (1991). *Loss aversion in riskless choice: A reference-dependent model*. In: «The Quarterly Journal of Economics», vol. CVI, n. 4, pp. 1039-1061.

WACHNER, J., ADRIAANSE, M., DE RIDDER, D. (2020). *The influence of nudge transparency on the experience of autonomy*. In: «Comprehensive Results in Social Psychology», doi: 10.1080/23743603.2020.1808782.

WASON, P., EVANS, J. (1974). *Dual processes in reasoning?*. In: «Cognition», vol. III, n. 2, pp. 141-154.