Università degli Studi di Urbino Carlo Bo

Department of

Pure and Applied Sciences

Ph.D. PROGRAMME IN: Research Methods in Science and Technology

CYCLE XXXVII

THESIS TITLE

Green Nudging, Environmental Beliefs and Self-Knowledge.

A Contribution to the Psychology of Nudging

ACADEMIC DISCIPLINE: PHIL-02/A

Thesis written with financial support from the
FSE-REACT-EU programme, PON "Ricerca e Innovazione" 2014-2020 (D.M. 1061/2021)
Azione IV.5 "Dottorati su tematiche Green"

Coordinator: Prof. Luca Lanci

Supervisor: Prof. Vincenzo Fano

Ph.D. student: Adriano Angelucci

ACADEMIC YEAR
2023/2024

Table of contents

Introduction

Human choices and behaviors have been at the heart of many human problems since the dawn of time, and humans have always been a singularly unruly and belligerent lot. As a consequence, they have long been in the business of exploring effective and efficient ways to intentionally steer people's behavior in a predictable direction without resorting to coercion. The age-long history of their endeavors on this front suggests that there are basically two broad ways in which one can pull that off. One can either intervene 'from the inside' on the various *beliefs* that often lead people to make corresponding decisions and choices, or one can approach the matter 'from the outside', by shaping the physical or now often digital *environments* in which those decisions and choices are made. In this work, I suggest to think of these two strategies respectively as the *internal*, and the *external* route to behavior change. As a powerful instrument in the hands of policymakers, the external route has gained considerable traction over the last few decades. Its fundamental insights can be traced back to research on human behavior and decision making initiated by economists and psychologists in the 1950s, and it currently travels under the banner of *nudging*.

The present dissertation is intended as a contribution to the psychology of nudging. It investigates both theoretically and empirically the as yet unclear and underexplored conditions under which people's beliefs can affect, and in turn be affected by, nudge-based policies. As such policies have so far proved extremely promising in the environmental domain, the investigation in question focuses in particular on pro-environmental beliefs and behaviors. The dissertation is premised on the methodological assumption that philosophical and psychological theories and results can often helpfully illuminate each other. Its guiding idea, in particular, is that attempts at developing a mature psychology of nudging can profit immensely from building on social psychology's long-standing efforts to clarify the many, and often counterintuitive, mutual interactions between the beliefs we hold and the behaviors we engage in. In this regard, two interconnected theses can be regarded as constituting the theoretical backbone of the whole dissertation. The first one holds that, to the extent that our beliefs have been amply shown to influence our behaviors, and can hance be often relied on to explain and predict them, one should expect that people's beliefs will have a sizeable impact on the nudging process. By parity of reasoning, the second thesis holds that, to the extent that our behaviors have been amply shown to influence our beliefs, and can likewise often be relied on to explain and predict them, it is also reasonable to expect that the nudging process will itself have a sizeable and measurable impact on the formation of people's beliefs. In line with its general goal, the dissertation consists of two parts – i.e. a theoretical, and an empirical one. Chapters 1 to 3 constitute its theoretical part. They provide a host of conceptual tools, and develop a series of considerations that are then relied on in order to put forward a specific, and empirically testable hypothesis concerning a possible consequence of nudging on the formation of our beliefs. Chapter 4 constitutes the empirical part. It presents and discusses the results of a pilot study designed and performed in order to begin testing the hypothesis in question. I what follows, I provide an overview of the four chapters.

Chapter 1 sets the stage for the considerations to be developed throughout the whole dissertation by proposing – in section 1.1 – that we think of our species along le the lines of an evolutionarily inspired picture. According to this suggestion, humans can be usefully regarded as 'natural born

influencers', the most powerful convincing machines that ever roamed this planet. Sections 1.2 and 1.3 then begin considering the ways in which natural born influencers have actually gone about the business of gaining compliance, and they are chiefly concerned with what I called the internal route to behavior change – i.e. persuasion. Section 1.2, in particular, details some of the main findings produced by early scientific research on the matter starting from the 1950s. As we shall see, the Yale group led by Carl Hovland had the lasting merit of unearthing a wealth of more or less intuitive persuasion-related phenomena, that I collectively refer to as the *persuasion data*. Section 1.3 then documents the emergence of dual-process approaches to human information processing, which, starting from the 1980s, allowed social psychologists to better account for the persuasion data by interpreting them in the light of more systematic theoretical frameworks. The section ends by discussing the two currently dominant dual-process approaches to attitude change – i.e. the *Heuristic-Systematic Model*, and the *Elaboration Likelihood Model*. Sections 1.4 and 1.5 then move on to the external route to behavior change – the one that this dissertation will be mainly concerned with. Section 1.4, in particular, introduces the notion of *nudging*, retraces its origins to economists and psychologists' early research on human judgment and decision making (i.e. the field we now refer to as *behavioral economics*), and discusses some of the central features of nudges by considering a famous example coming from the health domain. Section 1.5 focuses on the widespread reliance of nudge-based policies in the environmental domain, where they have proved extremely promising in fostering the adoption of what I call *climate-change-responsive behaviors*. It introduces and discusses the so-called *Value-Action Gap* – i.e. the well-documented, and long-investigated disconnect between our environmental knowledge and awareness, on the one hand, and our actual pro-environmental choices and behaviors, on the other. It also considers various psychological barriers that environmental psychologists regard as responsible for the existence of such gap. The section ends by considering various ways in which nudge-based policies can validly contribute to remove or at least reduce the effects of such barriers, and to therefore bridge the value-action gap by helping people realign their environmental choices with what they already appear to regard as the right thing to do.

Chapter 2 delves into the ethical dimension of the currently widespread reliance on nudge-based policies. It is premised on two interconnected methodological assumptions. The first one is that the psychology of nudging bears crucially on its ethics, as a fair assessment of the many ethical issued raised by the use of nudges in public policy, as it has been repeatedly observed, can only be reached through a preliminary solid understanding of the many different ways in which such tools can affect our everyday decisions and choices The second assumption is that it does not seem to make much sense to morally object to nudging *in general*, as different types of nudges will typically engage different psychological mechanisms, and should therefore be independently assessed. As I demonstrate by discussing and responding to a specific objection, once both of these assumptions are taken into due consideration, some of the arguments leveled against the use of nudging can be shown to lose much of their initially intuitive appeal, and hence of their force. Section 2.1 focuses on the fact that, although remarkably effective in fostering the adoption of CRBs, nudges have proved morally controversial, and their use in policy has been put under close ethical scrutiny. Many authors indeed worry that nudges may end up clashing with widely held human values, such as liberty, autonomy, respect, and dignity. Sections 2.2. and 2.3 engage closely with what, in my view, constitutes the most pressing objection to the use of nudging in public policy – i.e. the so-called *autonomy objection*, according to which nudges would represent a

serious threat to human rational agency. As I try to show, the autonomy objection ultimately rests on an implausible view of intellectual autonomy, coupled with a rather skewed understanding of nudging's cognitive underpinnings. Whereas section 2.2 expounds the objection by highlighting some of its more questionable epistemological assumptions, section 2.3. examines and defends a view of nudges recently developed in the work of Neil Levy, according to which nudges, once properly construes, would not only be wholly compatible with our rational agency, but also an extremely valuable way to more fully exercise it. In line with the second of the above-mentioned methodological assumptions, sections 2.4 and 2.5 then apply the abstract considerations so far developed to the concrete case of energy defaults. Section 2.4, in particular, discusses two field studies conducted in Germany, and testifying to the extraordinary power of defaults in decreasing the use of grey energy sources by steering households into electing greener energy tariffs. Section 2.5 engages closely with the psychology of defaults, and demonstrates that, once the documented metacognitive component of default-induced decisional processes is taken into proper account, energy defaults cease to constitute a threat to our rational agency.

From a general point of view, chapter 3 is intended, and hence best approached as just one long argument leading up to the central proposal of this dissertation, which is put forward in its last section. The theoretical backbone of this work, I said, consists of two interconnected theses. For the sake of brevity, we can rehearse them as follows: (1) to the extent that our beliefs influence our behaviors, the beliefs of the nudgee (the person being nudged) should have an impact on the nudging process; (2) to the extent that our behaviors influence our beliefs, the nudging process should likewise have an impact on the nudgee's beliefs. Approaching matters in a more analytical way, the chapter's structure can be seen as a breakdown of these two theses into their component parts. Section 3.1 centers on the idea that beliefs shape behaviors, and can hence be relied on to explain and predict them. In articulating this basic thought, the section retraces its grounding in a long-standing tradition of social psychological research on attitude-behavior consistency. This research tradition isolated a vast array of factors that are now known to moderate the relation between beliefs and behaviors, and to therefore affect the extent to which the former can reliably predict the latter – i.e. a plurality of findings that I selectively discuss, and collectively refer to as the *consistency data*. The section ends by expounding one of the two currently dominant theoretical frameworks for predicting behaviors from beliefs – i.e. Fishbein and Ajzen's *Reasoned Action approach*. To take center stage in section 3.2 is the opposite direction of the causal arrow – i.e. the idea that behaviors shape beliefs, and can hence be relied on to explain and predict them. In articulating this thought, the section retraces its grounding in an equally long-standing tradition of psychological research inaugurated by Festinger's *Theory of Cognitive Dissonance*. This research tradition isolated several factors that are now known to moderate the relation between behaviors and beliefs, and to therefore affect the extent to which the former can reliably predict the latter – i.e. a plurality of findings that I selectively discuss, and collectively refer to as the *dissonance data*. The section ends by introducing a particular way of accounting for the dissonance data – i.e. Bem's *Self-Perception Theory* – the main insight of which will play a key role in the last section. Section 3.3 discusses a long-acknowledged limitation of current nudge-based policies – i.e. the heavy context dependence of their effects – and illustrates it by means of two examples. Many today feel that a manifest replicability problem afflicts nudges' status as 'evidence-based policies'. What one indeed often observes is that a specific nudge which proved remarkably effective in one particular context, will prove largely ineffective (or worse, backfire) in another. Following

others, it is suggested that psychologists are best positioned to remedy this oft lamented state of affairs. The cure, it is argued, consists in opening nudges' black box and take as good a peek into it as we possibly can – i.e. reaching a deeper understanding of the cognitive mechanisms that underlie nudges' effectiveness. Section 3.4 then considers, discusses, and illustrates by means of several examples, some of the existing evidence for the proposition that the behaviors targeted by nudges, although often automatic in nature, will nonetheless in many cases be affected by the beliefs that the individual being nudged (the nudgee) happens to hold. Section 3.5, finally, discusses at some length the so far surprisingly neglected theoretical possibility that by directly affecting the nudgee's behaviors, a given nudge may also, under certain favorable conditions, indirectly affect her beliefs. In the course of doing so, the section introduces and explicates the notion of *nudged beliefs* – i.e. beliefs whose formation has been influenced by a nudge-induced action – which are then showed to constitute a subclass of a much broader family of mental states dubbed *epistemically engineered beliefs*. The section ends by developing a theoretical, yet empirically testable *self-interpretation model* of nudged beliefs. The model – which is inspired by Daryl Bem's Self-Perception Theory, and Peter Carruthers' *Interpretive Sensory-Access Theory* (ISA theory)– assigns a fundamental role to Self-Knowledge-related mechanisms in the formation of nudged beliefs, and builds on the fundamental insight that, on many everyday occasions, we more or less consciously self-attribute mental states based on an observation of our own past behavior, and the particular situation in which that behavior occurred.

Chapter 4 constitutes the empirical part of the present dissertation. It presents and discusses the results of a relatively simple pilot study designed and performed in order to begin testing the self-interpretation model of nudged beliefs put forward in chapter 3. The study targets the recycling behavior and environmental beliefs of a sample of first- and second-year undergraduate students enrolled at the University of Urbino. Leaving details aside, the model's central prediction is that performing a nudge-induced action will trigger a psychological process the output of which will be the self-attribution of a belief that the nudgee perceives as congruent with her own past action. As a consequence, the study aimed at assessing whether exposing individuals to a recycling nudge will have a positive impact on their self-reported environmental beliefs – i.e. whether individuals in the treatment group – who had been nudged to perform recycling action – will exhibit more pro-environmental (post-manipulation) beliefs on a currently standard scale. Statistical analysis revealed that the study did not find evidence for this effect. Section 4.4 discusses the results of the study, and, based on the existing literature, considers various possible explanations of why the treatment has not been found to make a statistically significant difference to the variable of interest. Its negative results notwithstanding, I take the study's main merit to consist in its having provided a useful illustration of how the philosophical and psychological ideas and considerations out of which my theoretical proposal was originally developed can readily be operationalized, and therefore, at least in principle, empirically validated. As a consequence, I believe that the study might constitute a valid starting point in order to design and perform future, and more elaborate experimental studies aimed at further investigating the effects of nudge-based policies on the formation of people's beliefs.

Acknowledgements:

Work on the present dissertation has greatly profited from numerous conversations with the following people, to all of whom I would like to express my heartfelt gratitude: Vincenzo Fano, Alberto Pirni, Vincenzo Crupi, Daniele Sgaravatti, Eugenio Orsi, Alessandro Chiessi, Stefano Calboli, Pierluigi Graziani, Christel Sirocchi, Niccolò Covoni, Matteo Bedetti, Matteo Antonelli, and Simone Smargiassi. I would also like to thank the University of Urbino for allowing me to make use of one of its beautiful buildings to conduct the empirical study reported in this dissertation, the many Professors who generously allowed me to take time out of their classes in order to recruit subjects amongst their students, as well as the students who kindly accepted to participate in the study. Special thanks go to Nicola Giannelli and James Newell, for having showed interest in the project and offered generous help in the recruitment phase; Manuela Berlingeri, for having offered valid advice concerning the experimental design of the study; Ariela Pagani for generously allowing me to attend her methodology classes, as well as for providing invaluable help with the questionnaire design and data gathering phase; Mirko Tagliaferri for having offered help with the online implementation, and administration of the questionnaire; and, last but not least, Matteo Perini, for his invaluable help with the data analysis.

1. Green Nudging

1.1. Natural born influencers

Nothing can make you feel more painfully inadequate than a loose-tongued fourth-grader. Clara, a ten-year-old digital native, was again reminding her increasingly uncomfortable grandmother of her lower status as a digital immigrant by rubbing the poor old lady's nose in her manifest social media illiteracy: "How can you not know who Chiara Ferragni is?" In some ways, it does not seem unlikely that future historians will look back at the currently enormous success of social media influencers as one of the defining features of early 21st Century's mass culture. At the same time, however, there can be little doubt that the set of psychological skills relying on which today's internet celebrities earn their living are not at all a result of the digital era. Indeed, once this phenomenon is approached from a broadly naturalistic perspective, it does not take much to realize that, although the tools of the trade have slightly changed, the overall aim of the game has nonetheless stayed largely the same: to intentionally bring it about, or make it more likely, that other individuals will think and act in a predictable way – i.e. to *gain compliance* without resorting to coercion.

In the wild, the ability to intentionally alter other individuals' behavior is clearly a valuable asset, as it is often what ultimately makes the difference between living to see another day, and starving to death or becoming the next predator's meal. Many species have therefore evolved specialized cognitive mechanisms aimed at deceiving.[1] Birds and mammals' sophisticated systems of alarm calls, for instance, can often serve this purpose. Vervet monkeys, to take just one example, rely on different types of calls to warn each other about the presence of various kinds of predators. Each call type will trigger a specific reaction – i.e. the one most appropriate to avoid the kind of threat at hand. If a snake approaches, the relevant type of call will cause other monkeys to climb *up* the nearest tree. If an eagle draws near, the relevant call will instead cause them to climb *down* that same tree in order to find shelter somewhere closer to the ground. Needless to say, this communication system comes in very handy when a monkey happens to spot a tree laden with juicy fruits that she wishes to feast on undisturbed, without having to share them with other members of its group. An apt strategy to achieve this goal, at least in the short-run, would be to emit an eagle signal when there are in fact no eagles in sight.[2]

What about humans? Well, on the one hand, we are hardly an exception to the general point of the above example. Our ancestors' remarkable proficiency at intentionally affecting the course of other individuals' actions – often, though not always, for self-serving purposes – is arguably one of the main reasons why we have ancestors at all. Just as other species, our own kind as well has long been in the business of influencing behavior. It must be acknowledged, however, that we are animals with highly developed cognitive skills – we possess uniquely human abilities that allow us to greatly outperform other animals at this game. The most obvious one is *language*. Human language differs markedly from any other communication system found in nature, however

[1] Cf. e.g. Trivers 2011, Ch. 2.

[2] The example comes from Mercier 2020, who draws on Seyfarth et *al.* 1980. Cf. Mercier (2020: 20).

sophisticated. So markedly, indeed, as to make for a sort of often unnoticed superpower when it comes to shaping other people's mental states. Steven Pinker, a leading cognitive psychologist, puts it as follows:

> "You and I belong to a species with a remarkable ability: we can shape events in each other's brains with exquisite precision … Simply by making noises with our mouths, we can reliably cause precise new combinations of ideas to arise in each other's minds."[3]

Those "combination of ideas", as we shall see in more detail in chapter 3, often lead to behaviors, and this is what makes language an extremely powerful compliance-gaining tool. A notable fact, in this regard, is that the evolution of language in our species is commonly thought by scholars to be deeply intertwined with the evolution of what psychologists refer to as *mindreading* – i.e. the ability to mentally represent and reason about other individuals' mental states, such as beliefs, desires, intentions, emotions or thoughts.[4] Although we constantly rely on this ability to navigate our social environment, mindreading, just as language use, comes so natural to most of us as to go almost unnoticed. Research has shown, however, that a severe impairment of this faculty would prevent us from making full sense of other peoples' words and actions, thereby seriously disrupting our most basic social interactions.[5] Starting at an early age, we become adept at the art of ascribing desires and beliefs to others in order to explain and predict their behavior, and – needless to say – the better we get at anticipating each other's actions, the more chances we have to successfully manipulate them. Skilled mindreaders make effective influencers.

It would appear, however, that in considering ways in which we differ from other animals with respect to our ability to influence other minds, we have so far been neglecting the real elephant in the room – i.e. *reason*. False modesty aside, the careful study of other animals' cognition, while extremely useful and arguably fundamental to fully understand our own, delivers an unambiguous verdict: we are clearly the smartest guys around here.[6] This is why our species is called *homo sapiens*. We possess an unmatched ability to solve novel problems, which led to the development of science and technology. We also display the highest level of behavioral plasticity. Contrary to any other species, we can easily acquire new skills, and therefore adapt to very different environments. This allowed us to spread throughout the world and colonize even the most inhospitable regions of this planet. It is no wonder, then, if the Western philosophical tradition, in its struggle to capture the essence of human nature, singled out reason as the mental faculty that distinguishes us from other animals – reason is what makes us special! If so, couldn't it also be what ultimately makes us excel at compliance gaining? A very interesting hypothesis, in this regard, has been put forward by Hugo Mercier and Dan Sperber.[7]

According to the received, and still largely predominant view on the matter, it would be fairly obvious that the primary function of human reason is to enhance individual cognition – i.e. to

[3] Pinker (1994: 13).
[4] Cf., e.g., Apperly (2010).
[5] Cf., e.g., Baron-Cohen (1997).
[6] Cf. e.g., Anderson 2010.
[7] Mercier & Sperber (2017).

form better beliefs and make better decisions.[8] In their book *The Enigma of Reason*, Mercier and Sperber have challenged this centuries-old assumption – which they regard as an empirically ill-supported dogma – by extensively arguing for the intriguing idea that this ability in fact evolved primarily for social purposes – i.e. to facilitate social interactions. On the *interactionist approach* they advocate, reason would be best seen as a social competence, a biological adaptation to the very special ecological niche that we happen to inhabit. Humans differ from other animals in the way and extent to which they cooperate and communicate. As a consequence, our social interactions pose unique problems of coordination and trust, problems that our reasoning ability, in their view, evolved to solve. Its two main functions, in particular, would be to *justify* one's beliefs and actions in the eyes of others, and to produce arguments intended to *convince* others to think and act as we suggest. Although the details of Mercier and Sperber's so-called *argumentative theory of reasoning* are rather complex, and have therefore been subject to sustained discussion and criticism[9], what matters for our present purposes is rather the general picture of our species that emerges from the overall argument developed in their book.[10] That picture could be summarized in a slogan by saying that humans, according to Mercier and Sperber, are natural born influencers – i.e. they evolved under a pressure to gain compliance. This idea makes for a convenient starting point for developing the considerations that follow. To put it emphatically: we are the ultimate convincing machines, the most powerful manipulators that ever roamed this planet.

It is then time to take a closer look at the specific ways in which humans have actually gone about the business of gaining compliance, and, in so doing, we could take the cue from Pinker's words, quoted above. One thing that our species is very good at, as he puts it, is shaping events in each other's brains by simply making noises with our mouths. The noises in question, of course, are words, and at least some of the events in question, as we shall see, are beliefs, or better the various psychological processes that eventually lead to their formation. As a matter of historical fact, the main strategy adopted by humans to influence each other's behavior – i.e. the way in which they *act* – has traditionally consisted in attempting to alter the way in which they *think* and *feel* about things through verbal communication – i.e. by indirectly working on their currently held beliefs and attitudes, under the commonsensical assumption that these mental states will at some point eventually lead to corresponding behaviors. In the next two sections, we will therefore selectively consider some of the main findings and explanatory frameworks produced and developed by contemporary research on persuasion.

---

[8] Mercier and Sperber refer to this stance as the *intellectualist approach* to reason, which they take to contrast sharply with the main tenets and predictions of their own *interactionist approach*. By their own admission, the approach was inspired by the program of "evolutionary psychology" outlined in the 1980s by psychologist Leda Cosmides and anthropologist John Tooby (Cf., e.g. Mercier & Sperber 2017: 181). Although Cosmides and Tooby's approach to the study of human behavior has proved controversial, assessing its overall viability is clearly beyond the scope of the present discussion. For an introduction to this approach, see the papers collected in Barkow, Cosmides and Tooby 1992. For an overview of the main lines of criticism, see Downes 2024.

[9] Cf., e.g., Mercier & Sperber 2011.

[10] A problematic aspect of their overall argument, in my view, is that the distinction between 'reason' in the empirical sense of 'reasoning ability', and 'reason' in the normative sense is often not as clear-cut as one would wish.

1.2. Who says what to whom

The beginning of the UNESCO Constitution, signed by 37 countries on 16 November 1945, reads as follows:

> "Since *wars begin in the minds of men*, it is in the minds of men that the defences of peace must be constructed."

The drafters of this document had little doubt about the fact that beliefs have consequences. Over the past few decades, a whole generation of human beings had come to believe that war was unavoidable. That belief brought about 80 million deaths – the deadliest conflict in recorded human history. It was indeed high time to collectively acknowledge that "wars begin in the minds of men", as well as to consequently address the social and psychological roots of that devastating carnage. The systematic study of persuasion in modern social psychology originated is this gruesome context.

In the U.S., worries about the effects of the propaganda used by Fascist and Nazi regimes in run-up to World War II had created an urgent practical need to reach a deeper understanding of the many ways in which peoples' beliefs can be changed through communication. What leads people to change their minds? Under which conditions are they most likely to question and eventually modify their extant opinions in response to a message aimed at persuading? Carl Hovland, a young psychologist from Yale, was appointed to lead the Information and Education Division created within the U.S. War Department with the mandate to put together a research group and investigate these questions in order to enhance soldiers' attitudes toward the war. Hovland and his team – which later became known as the Yale group – first set out to systematically codify the principles of persuasion by experimentally manipulating the factors (independent variables) which were at the time thought to increase or decrease the effectiveness of a communication, and measuring their effects.[11] Although the Yale group, partly due to the nature of their mission, was less interested in theorizing about the psychological mechanisms ultimately responsible for bringing about the observed effects, it nonetheless had the lasting merit of unearthing a wealth of more or less intuitive persuasion-related phenomena – henceforth collectively referred to as the *persuasion data* – that later social psychologists then attempted to more fully explain over the following decades. Most importantly, the group's investigations were structured according to an organizational principle that proved very influential, and informed all subsequent research on persuasion.[12] The independent variables assumed to affect the effectiveness of communication were indeed grouped under three main categories, pertaining to the speaker (*source* factors), the communication itself (*message* factors), and the audience (*recipient* factors) respectively. Political scientist and communication theorist H. Lasswell famously summarized these three components in the slogan: "Who says what to whom … and with what effect?"[13]

---

[11] Cf. in particular, Hovland et *al.* (1949) and Hovland, et *al.* (1953).

[12] The principle in question is usually traced back to Aristotle. In his *Rhetoric*, the Greek philosopher famously observed that: "Of the modes of persuasion furnished by the spoken word, there are three kinds. The first depends on the personal character of the speaker, the second on putting the audience into a certain frame of mind; the third on the proof, or apparent proof, provided by the words of the speech itself" (*Rhetoric*, Book I, ch. 2, 1356a, translated by W. Rhys Roberts).

[13] Lasswell (1948).

Let us then take a closer look at some of the persuasion data documented by the Yale group. Whereas some of the finding produced by Hovland's team, as we shall see, are more or less in line with our commonsense expectations, some others, on the contrary, proved puzzling and surprising. This does not mean, of course, that the former are less important of the latter to the study of attitude change. Indeed, as it is well known, casual observation is far from being a reliable guide when it comes to detecting the covariation (let alone the degree thereof) among events, especially if the events in question are not temporally close to each other, but separated in time. Our lay intuitions about relations among events, in particular, are notoriously prone to 'detect' all kinds of illusory correlations, often with disastrous consequences. In this regard, the following passage from Richard Nisbett, a prominent social psychologist, is worth quoting in full:

> "When we try to assess the correlation between two events that are *plausibly* related to each other – for which we are prepared to find a positive correlation – we're likely to believe there is such a correlation even when there isn't. When the events aren't *plausibly* related, we are likely to fail to see a positive correlation even when a relatively strong one exists. Worse – we're capable of concluding there is a positive relationship when the real relationship is negative and capable of concluding there is a negative relationship when the real relationship is positive."[14]

With respect to the study of attitude change, then, it is one thing to find the existence of given persuasion phenomenon commonsensical or intuitive, and quite another to possess at least some experimental evidence in favor of its actual existence.

Starting with source factors, research shows that, as Aristotle had anticipated, attitude change can be remarkably affected by certain characteristics of the person who delivers the message – i.e. the communicator – in spite of identical message contents. The persuasiveness of a communication, in particular, seems to largely depend on the overall *credibility* of the speaker. Credibility is in turn standardly construed as a function of both *expertise* – i.e. the speaker is knowledgeable about the issue at hand – and *trustworthiness* – i.e. she intends to deliver correct information. It is important to note, however, that both characteristics are here typically understood not as actual, but only as perceived ones.[15] To exemplify, a speech about the imminent dangers of anthropogenic climate change will be more likely to shift our opinions on the matter if coming from the mouth of a climate scientist, than if coming from that of a barber – i.e. perceived high-expertise tends to produce more attitude change.[16] Perceived trustworthiness, on the other hand, can be enhanced, for instance, if the speaker defends a position that appears to go blatantly against her own self-interest – think, e.g., of a pro-environmental speech delivered by Donald Trump.[17]

Just as one would have expected, then, there is clear evidence that, in general, low credibility leads to less persuasion, presumably because both a perceived lack of expertise on the part of the speaker or her perceived untrustworthiness tend to work as discounting cues in the eyes of the audience – i.e. signals that the recipient would be better off not giving too much weight to the

---

[14] Nisbett (2016: 145, my emphases).

[15] Hovland and colleagues indeed define 'expertise' as the "extent to which a communicator is *perceived to be* a source of valid information", and 'trustworthiness' as the "*degree of confidence* … in the communicator's intent to communicate assertions that he considers most valid." (Hovland et al. 1953, my emphases).

[16] Hovland and Weiss (1951).

[17] Eagly et *al.* (1978).

message or else completely disregarding it. All in all, this does seem in line with lay intuition. Yet, one might wonder, is it *always* the case that low credibility produces less persuasion? In this regard, further investigations of the effects of source credibility on attitude change soon led to one of the most surprising findings in persuasion research. A central question on the Yale group's agenda concerned time. Attitudes in general are known to decay over time. How about attitudes changed in response to a persuasive message? How long do persuasion effects last? In a seminal study on the matter, Hovland and colleagues made an astonishing discovery: a few weeks after being exposed to a message, the persuasion created by a high-credibility speaker tends to decline, whereas the persuasion created by a low-credibility speaker tends to increase – i.e. over time, low credibility leads to more persuasion, not less, which means that, as time goes by, low-credibility and high-credibility speakers end up being just as effective![18] This puzzling phenomenon has since been known as the *sleeper effect*, and it has spawned a long series of studies aimed at clarifying its underlying mechanisms.[19]

Similar counterintuitive findings come from research on message factors – i.e. the many aspects of the persuasive message itself that can render it more or less effective. One such aspect is *fear*, which happens to constitute one of the most powerful drives of human (just as other animals) behavior. There is indeed little doubt that fear can motivate us to alter our behaviors in order to avoid the unwelcome consequences of our actions. It is no wonder, then, if persuasive messages aimed at gaining support for social policies often appeal to this primal emotion. Think of the gruesome pictures of rotten lungs or hospitalized end-of-life patients (and captioned with explicit messages such as 'smoking kills') printed on cigarette packs by anti-smoking campaigners or else of the apocalyptic images of natural disasters (often coupled with equally threatening messages, such as 'there is no planet B') currently circulating on the media to increase people awareness of the dangers posed by anthropogenic climate change. There is however a question of *how much* fear is likely to be most effective in bringing about behavioral change? An inference that one might be tempted to draw, in this regard, is that more fear leads to more change. Early research on pro-health behavior, however, indicated that, contrary to what one would expect, high levels of fear are counterproductive for persuasion.[20] In fact, the psychology of fear appeals seems even more complex. Interestingly, subsequent research strongly suggested that the effectiveness of such appeals is a function of subjects' perceived self-efficacy – i.e. that persuasive messages appealing to fear are only effective to the extent that they also provide their audience with a clear indication that, by acting (or avoiding to act) in specified ways, they can avoid the ominous consequences of their relevant actions.[21] It would appear, that is, that we are not only natural born influencers, but also natural born Stoics – why worry about something, if there isn't anything I can do about it? Needless to say, this has immediate implications for environmental campaigns. In the case of the latter, indeed, the take home lesson seems to be that there is little hope of changing peoples' everyday climate-impacting behaviors unless our messages are coupled with clear statements that, yes, there *is* something we can all do about it.

Speaking of clarity, a natural thought is that clearly stated messages will be more likely to capture the audience attention. *Attention*, indeed, is another aspect investigated by research on message

[18] Hovland et *al.* (1949), Hovland & Weiss (1951).

[19] Kumakale & Albarracín (2004), a meta-analysis based on 72 studies of this phenomenon, confirms the existence of the sleeper effect.

[20] Janis & Feshbach (1953).

[21] Cf., e.g., Witte & Allen (2000) for the case of public health campaigns.

factors. The central question here is as follows: How does the amount of attention that a recipient gives to the content of a persuasive message affect attitude change? The intuitive inference one might be inclined to draw, with respect to this issue, is that the more carefully one attends to a message, the more likely she will be to be persuaded by its content. As it turns out, however, this is not always the case. Early research indeed showed that sometimes messages including elements intended to distract the audience – i.e. to divert their otherwise undivided attention from the message content – are the one that prove most effective. Although counterintuitive, then, more distraction can in effect, under certain conditions, actually lead to more attitude change. What are the conditions in question? In this regard, much seems to depend on whether, and to what extent, the recipient of a message agrees with the particular position advocated by the speaker. Indeed, what has been repeatedly observed is that while listening to a message that we happen to disagree with, we automatically begin to look for counterarguments in our heads – i.e. we initiate a search for countervailing reasons that could plausibly be appealed to in order to resist the message. As a consequence, at least in the case of recipients who are likely to disagree with a given persuasive message, coupling the latter with distracting stimuli has the effect of disrupting her mental search – i.e. to inhibit her ability to generate counterargument, and this interference in turn leads to more attitude change. As advertising professionals know very well, it is not always easy to find cogent arguments that will convince prospective customers to waste their money on a product that they do not wish, nor need to possess. In such circumstances, commercials that distract them with irrelevant details (e.g. an attractive model or a humorous line) prove to be very effective.[22]

As I mentioned above, the Yale group was chiefly interested in systematically documenting the *factors* that affect the effectiveness of a message, and less interested in theorizing about the *processes* that actually lead to attitude change. In this regard, it must be noted that the dominant approach to cognition at the time were various learning theories based on Pavlov's classical conditioning, Skinner's operant conditioning, and Hull's drive theory.[23]In line with this tradition, Hovland and colleagues typically construed persuasion as a matter of *learning*, thereby suggesting that appealing to this general concept would be sufficient to account for at least the bulk of persuasion data unearthed by their studies.[24] A central assumption, in particular, was that persuasion attempts, being a function of learning, would only be effective to the extent that they facilitated message learning. As it is evident in the case of the distraction effects discussed above, however, some of the factors that are observed to affect attitude change do not seem to lead to increased message learning. Such anomalies in the data – as it is frequently the case in science – were either not noticed or else conveniently swept under the carpet. Thankfully, however, there is only so much dirt that you can hide under a single rag – in this case, the learning model. Indeed, as anomalies kept accumulating, it became progressively clear that – since learning can apparently occur in the absence of attitude change, and vice versa – message learning could not be the only process at work in persuasion.[25]

An important step forward with respect to the learning model – which partly reflects the passage from the behaviorist to the new cognitive era – was represented by the *cognitive response model* of

[22] Festinger & Macoby (1964), Osterhouse & Brock (1970), Strick et *al.* (2012).
[23] Cf. Pavlov (1927), Skinner (1938), and Hull (1943) respectively.
[24] Cf., e.g., Hovland & Mandell (1952), and Kelman & Hovland (1953).
[25] Cf. Briñol & Petty (2009).

persuasion, first proposed by Anthony Greenwald in the late 1960s.[26] Persuasion, according to this approach, is not just a matter of understanding and remembering the *content* of a message. In fact, what matters the most is our *cognitive response* to that content – i.e. the mixed bag of thoughts and feelings that are automatically activated while listening to the message, and which can of course vary both in valence and intensity. Quite often, indeed, we seem far better able to remember our responses to a message than the message itself. Think of the following two common experiences. While thinking back on a picnic with your partner, you recall that he said something that crossed you, yet you can no longer remember what it was – i.e. you do not remember his words, but only how they made you feel. While telling a friend about a conference that you recently attended, you recall that the speaker, in order to support her views, appealed to an argument that you found unconvincing, yet you no longer precisely remember how the argument went. What ultimately drives persuasion, and to what extent, according to Greenwald, would not be the content of a message. We are not directly moved by words, but rather, indirectly, by the way in which they make us feel as well as by the thoughts they trigger. Moreover, the valence and intensity of our cognitive responses to a persuasive argument do not necessarily depend on its strength, which means that the same argument can at times have opposite effects on different recipients.

While certainly constituting a progress with respect to the learning approach, however, even the cognitive response model ultimately proved insufficient to fully account for the persuasion data. Most importantly for our purposes, the idea began to spread that searching for a *single* cognitive process – be it learning or cognitive responses – underlying all persuasion phenomena was a dead end. The main reason, according to many, was the human mind does not have just one, but *two* different ways of processing information, which are activated under different circumstances and eventually lead to attitude change in response to a persuasion attempt. It is to this intriguing idea that we shall now turn.

1.3. It takes two processes to tango

Here is a puzzling fact about humans. On the one hand, we are hopeless control freaks. We like, perhaps even need, to believe that both our minds and our environment, other people included, are a lot more under our voluntary control than it is actually and demonstrably the case – often with disastrous consequences.[27] At the same time, however, the more or less inchoate suspect that there may in fact be a lot more going on in our heads than it appears in or is accessible to our consciousness mental life, and the corresponding worry that most of our thoughts, decisions, and actions, unbeknownst to us, are massively affected by this hidden mental activity have been around for centuries.[28] In the Western tradition, most theoretical work aimed at better articulating this widespread intuition – from Plato to Freud – usually did so based on informal observations of human behavior coupled with personal introspection. This changed with the emergence of

[26] Greenwald (1968). See Petty & Brock (1981) for a review.

[27] Cf., e.g., Taleb (2010).

[28] For a survey of the Western history of thinking about mental duality, see Frankish & Evans (2009). For a cross-cultural approach to the same issue, see Buchtel & Norenzayan (2009). For a thought-provoking attempt to show that the notion of unconscious processing is completely off, and that there is really nothing more going on in our minds than what appears in consciousness – i.e. that the mind 'is flat' – see Chater (2018).

experimental psychology as a discipline, as early evidence of unconscious mental processing began to slowly accumulate. During the 1960s and 1970s, the cognitive revolution in psychology, and the widespread acceptance of a computational approach to the mind, led to the explanatorily powerful notion of a *cognitive unconscious* – i.e. a complex form of information processing that goes on below the threshold of our awareness.[29] Social psychologists, in particular, began to carefully investigate the now well supported idea that, in spite of our persistent intuitions to the contrary, our behaviours are in fact controlled by a host of unconscious processes, and that we abundantly rely on conscious reasoning to confabulate explanations for those behaviors.[30]

It is against this background that the idea according to which the human mind would have not one but *two* different ways of processing information first came into clear focus as an empirically testable hypothesis, and evidence today strongly suggests that this hypothesis is at least broadly correct. During the 1970s, a number of researchers studying different aspects of human higher cognition independently converged on this insight, and so-called *dual-process theories* began to sprout within several, although in practice largely disconnected, domains of human psychology – in particular deductive reasoning, decision making, and social judgment.[31] What all these proposals had in common was their positing of two distinct types of mechanisms that the mind can apply to one and the same cognitive task – processes that rely on fundamentally different procedures and that, as a consequence, can often lead to conflicting results. Following current use, we will refer to them as System 1 (S1), and System 2 (S2) processes. Whereas S1 processes are typically characterized as fast, automatic, and unconscious, S2 processes are characterized as slow, deliberative, and conscious.[32] Authors however currently disagree on how the two types of processes interact. In this regard, Evans (2008) has suggested that we can usefully divide dual-process theories into two broad camps, which he dubs *default-interventionist* and *parallel-competitive* theories respectively.[33] Authors belonging to the former camp, as its name suggests, would share the assumption that S1 processes supply contents for the conscious reasoning carried out by S2. S1, in particular, would prompt behaviors that S2 will then either accept or inhibit. Authors belonging to the latter camp, would instead share the assumption that S1 and S2 processes stem from two parallel forms of learning, which in turn lead to two forms of knowledge – i.e. implicit and explicit knowledge – constantly competing to control behavior. Let us first take a quick look at dual-process theories in the domains of reasoning and decision making.

The spread of dual-process ideas in the psychology of deductive reasoning corresponded to researchers growing interest in understanding how contextual factors, including a subject's prior beliefs and knowledge, affected performance on various reasoning tasks such as, e.g., the famous four-card selection task invented by Peter Wason.[34] A within-person conflict between contextual factors and normative standards emerged perhaps most clearly in the study of what has come to

[29] Cf, in particular, Kihlstrom (1987) and Reber (1993).

[30] Cf., e.g., Nisbett & Wilson (1977); Wilson (2002) and Wilson & Dunn (2004).

[31] Although such proposals were initially largely independent from one another, some theorists subsequently attempted to unify them into broader theories of mental architecture. Their pioneering attempts were guided by the idea that there may indeed exist two architecturally and evolutionarily distinct cognitive systems undergirding the two types of processing. Cf., e.g. Evans & Over (1996); Sloman (1996) and Stanovich (1999, 2004).

[32] As it is to be expected, given the complexity of the matter, things are not nearly as clear cut, especially with respect to System 1 processes. Cf. Evans (2008). This rough characterization will however still do for our present expository purposes.

[33] Cf. Evans 2008: 266, 271.

[34] Cf., e.g., Wason & Evans (1975).

be known as *belief bias*.[35] In the standard paradigm, subjects are asked to assess the logical validity of syllogisms whose validity, and the believability of whose conclusions, has been experimentally manipulated – i.e. whereas some syllogism are logic-belief compatible, some others are instead intentionally designed to generate conflict by featuring either valid arguments with unbelievable conclusions or invalid arguments with believable conclusions. What researchers repeatedly found are a main effect of logic (conclusions of valid syllogisms are accepted more often), a main effect of belief – the belief bias – (believable conclusions are accepted more often), and an interaction effect between belief and logic (the belief biased is much stronger for invalid syllogisms). These findings have been replicated many times. As the shift from logical to belief-based reasoning is a function of factors such as, e.g., time pressure and working memory overload, they are typically taken to constitute evidence in favor of dual-processing.[36] Similar dual-process approached were later applied to the study of judgment and decision making as well. In the early 2000s, for instance, Kahneman and Frederick proposed a theory of probability judgment, according to which – in line with the above-mentioned default-interventionist picture – S1 heuristic (and often biased) judgments would provide intuitions to the S2 analytic reasoning, whose main function would be to either accept, inhibit or improve them.[37] A limit of these dual-process approaches to reasoning and decision making is that while S2 processes, although admittedly not infallible, are nonetheless generally associated with normative correct responding, S1 processes are typically associated with biases. This generates the misleading impression that S2 processes are always to be preferred over S1 ones. In the case of decision making, for instance, S1 processes often seem to play a key role in the expert judgments and decisions of fire officers and paramedics[38], which suggests that, in general, conscious, deliberative, S2 reasoning can at times impair good decision making.[39]

In light of the above, let us now go back to persuasion as a means of compliance gaining. Dual-process approaches became the dominant paradigm in social psychology during the 1980s, and they have since been consistently applied to investigate the automatic and unconscious processing of information in various domains of social cognition, such as person perception, stereotyping and – crucially for our purposes – attitude change.[40] As we saw in section 1.2, early research on this latter topic eventually proved insufficient to fully account for the wealth of persuasion data unearthed by Hovland's group, as its most influential explanatory frameworks – i.e. the learning model, and the cognitive response model – assumed that all persuasion phenomena should be governed by *one* and the same cognitive process. This underlying assumption, however, was hardly compatible with the observation that a number of variables – such as, e.g., the credibility of a speaker – prove highly effective in some persuasive context, but almost ineffective in others. A turning point, in this regard, was represented by the simultaneous and independent development of two dual-process approaches to attitude change: the *Heuristic-Systematic Model* (HSM), and the *Elaboration Likelihood Model* (ELM).[41] The reason why, e.g., a speaker's credibility is likely to have a strong impact in some contexts, yet to be virtually inconsequential in others,

[35] Cf. Evans et *al.* (1983).
[36] Cf., e.g., Klauer et *al.* (2000).
[37] Cf. Kahneman & Frederick (2002, 2005).
[38] Cf. Klein (1999).
[39] Cf. Wilson & Schooler (1991).
[40] Cf., e.g., Chaiken & Trope (1999); Wilson (2002): Bargh (2006).
[41] For the HSM, see Chaiken (1980, 1987), and Chaiken & Ledgerwood (2012). For the ELM, see Petty & Cacioppo (1981, 1984), and Petty & Briñol (2012).

according to both HSM and ELM, is that, in general, people rely on *two* fundamentally distinct ways of processing information when they are exposed to a persuasive message. Both models, in particular, posit that which process is activated will be determined by two key factors – i.e. our *motivation* to process the message, as well as our *ability* to do so – and both of these aspects vary widely across persons and situations. At a practical level, this is bound to make life harder for the compliance professional, as deciding which type of message will be most likely to, e.g., convince people to support a given environmental policy, is something that cannot be done in the abstract, but rather calls for a careful analysis of which type of process the message will more likely trigger across people and situations. Let us then take a quick look at both HSM and ELM.

The HSM distinguishes between *heuristic* and *systematic* processing. The former constitutes our default processing mode, it does not require effort, and it is automatic and associative in nature. It does not engage in an analytic elaboration of incoming persuasive messages, but rather relies on quick and dirty 'rules of thumb'. A recipient, for instance, may be persuaded by a speech about the imminent dangers of anthropogenic climate change, not as a consequence of a careful scrutiny of the arguments presented, but rather because the position advocated by the speaker happens to represent the majority position, and, in the recipient's past experience, majorities have usually turned out to be correct. *Systematic* processing, on the other hand, is effortful and involves deep thinking and careful consideration of the argument presented in the message. When activated, it occurs in addition to and simultaneously with the default, heuristic mode. As it is very demanding of cognitive resources, it is driven by a 'sufficiency criterion'. In other words, it constantly aims at striking an optimal balance (varying across dispositional and situational factors) between the motivation to hold correct attitudes and process information objectively, and the mental effort required to do so. In a similar fashion, the ELM draws a parallel distinction between *central* and *peripheral* processing, construed as two different routes to persuasion and attitude change. These two processes, however, while qualitatively distinct, are posited by the ELM to represent the two end points of one and the same *elaboration continuum*. Conditions of high motivation and ability to process a message (typically triggered by high personal involvement with its content) lead to the activation of central processing, and therefore lie on the high-elaboration-likelihood end of the spectrum; whereas conditions of low motivation and ability lead to the activation of peripheral processing, and hence lie on the low-elaboration-likelihood end. An interesting finding, in this regard, concerns the relationship between number and strength of arguments under different elaboration conditions. It appears, indeed, that whereas in conditions of high elaboration 9 weak arguments lead to *less* persuasion than 3 weak arguments, in conditions of low elaboration the pattern reverses, and 9 weak arguments lead to *more* persuasion than 3 weak ones.[42] Whereas central processing typically assesses arguments' quality in terms of their strength, peripheral processing assesses it in terms of their number!

The predictions of both HSM and ELM have so far been tested through hundreds of studies, all sharing the same experimental paradigm, involving the manipulation of three basic independent variables: systematic *vs.* central cues contained in the message (such as, e.g., argument strength), heuristic *vs.* peripheral cues contained in it (such as, e.g., source attractiveness), and the recipient's motivation or ability to process (usually by exploiting personal relevance, distraction or cognitive load imposed by a secondary task).[43] Although there are notable differences between the two

[42] Petty & Cacioppo (1984).
[43] Petty et *al.* (2003).

models, let us conclude this section by highlighting what they have in common. To begin with, both models posit that attitudes formed by relying on the more effortful (systematic/central) type of process will be likely to prove more persistent over time, resistant to counterexamples, and – importantly for our purposes – predictive of behavior. They moreover assume that the two types of processes can often run in parallel, and – differently from most dual-process approaches to deductive reasoning and decision making – they explicitly acknowledge that both processes, not just the heuristic/peripheral one, can give rise to biased judgments. A final point well worth mentioning is that both HSM and ELM postulate that, although humans are generally moved by a default motivation to hold correct opinions, they also care a great deal about other goals beside knowledge – i.e. they are also moved by other motivations, such as a desire to protect their current beliefs or longing for social acceptance. Such goals, while not in the service of accuracy, can nonetheless still powerfully affect the processing of persuasive messages.

1.4. A new policy game in town

The considerations so far developed in this chapter started from elaborating on an evolutionarily inspired picture of our species. Taking the cue from Mercier and Sperber's groundbreaking work on the social functions of our reasoning ability, I suggested to think of humans as 'natural born influencers'. On this picture, we evolved under a powerful social pressure to gain compliance – i.e. intentionally make it the case, or make it more likely, that other individuals will behave in a predictable way, without necessarily resorting to coercion. A long-practiced way to achieve this goal, as we have seen in the last two sections, has traditionally consisted in intervening on people's beliefs and attitudes by means of verbal communication. Since its early days, the scientific study of persuasion has indeed been premised on the implicit assumption that altering the way in which people think and feel about various issues will eventually bring about a change in the way they act.[44] Persuasion, then, is one way to influence people's behavior. The rub, however, is that as every parent can witness on a daily basis, as the scant effects of social and political campaigns stand to show, and as psychological research amply confirms, trying to convince humans to behave or refrain from behaving in a certain way by means of explicit arguing usually proves a very tough nut to crack.[45]

As social psychologists have known for a long time, this is not just due to the fact that some persuasive messages are admittedly poorly crafted, but also to the further fact that people are often strongly motivated to defend their current attitudes (especially when the latter are linked to their personal values) by actively resisting persuasion in various way, such as, e.g., counterarguing or derogating the source of a message.[46] Tilting people's opinions in a desired direction, in short, is demonstrably much harder than one would casually assume. Moreover, even in those cases where persuasive attempts *have* been successful, and attitude change has indeed occurred, a long tradition of research on attitude-behavior consistency has repeatedly shown that explicit attitudes

[44] It is perhaps worth noting, in this regard, that Hovland and his colleagues, in line with the spirit of their times, typically referred to attitudes as 'implicit behaviors' (cf., e.g. Hovland et *al.* 1957).

[45] Cf. Mercier (2020) for a recent, evolutionary minded, extended defense of the claim that people are in fact much harder to be swayed by persuasion, and much less gullible than it has been traditionally assumed in social science.

[46] Cf., e.g., McGuire (1961, 1964), Brehm (1966), and Brehm & Brehm (1981).

do not always lead to corresponding behaviors.[47] It follows that, in general, successfully shifting the particular way in which someone thinks and feels about an issue is usually not *sufficient*, by itself, to make it more likely that she will act in a way congruent with her newly acquired position. The good news, for our present purposes, is that a growing body of evidence coming from research on human judgment and decision-making indicates that directly intervening on people's beliefs and attitudes is not always *necessary* to affect the way in which they act – i.e. the behaviors they adopt, as well as the decisions and choices they make. One could put this by saying that there are in fact both an *internal*, and an *external* route to behavior change. The internal route aims, as it were, at altering behavior 'from the inside' – i.e. by intervening on the *beliefs* and *attitudes* that (sometimes) lead people to make corresponding decisions and choices. The external route, on the other hand, aims at achieving this same goal 'from the outside' – i.e. by intervening on the physical or digital *environment* in which those decisions and choices are made. Within the social sciences, this latter route has gained considerable traction over the last several decades. Although its fundamental insights can be traced back to research on human behavior and decision making initiated by economists and psychologists in the 1950s, one of its most recent implementations, as we shall see, currently travels under the banner of *nudging*.

The concept of nudging originated in behavioral economics, a relatively new and thriving field of economics that led Nisbett to half-jokingly rename the latter discipline 'the *formerly* dismal science'.[48] What we now refer to as behavioral economics gradually developed as a reaction to a set of general assumptions concerning human judgment and decision making underlying standard economics. A long process, initiated in the late 18th Century, had indeed culminated in the mid-20th Century with the formalization of a number of axioms that economic agents ought to obey when making decisions under conditions of risk or uncertainty in order to maximize utility.[49] This set of formal rules – known as *expected utility theory* or else the *standard model of rational choice* – still constitutes today the dominant normative theory of decision making under risk or uncertainty in standard economics. Starting from the 1950s, some maverick economists began to take issue with the overall picture of human rationality that emerged from the standard model. Real-life agents, they felt, could hardly be expected to live up to the standards of rationality prescribed by expected utility theory, and therefore built into standard economic models.[50] To be fair, at least in theory, the standard model had not been conceived as a descriptive or predictive account of human behavior. In practice, however, its widespread adoption had nonetheless resulted in what has been recently portrayed as an ambiguous "blurring of the lines between normative and descriptive assumptions in economics".[51] The fundamental merit of early behavioral economists consisted in pressuring mainstream economists into coming out in the open about this ambiguity, and to systematically address the many observed 'anomalies' in human judgment and decision making, instead of chocking them up to 'random errors'. Initially this attempt proved largely ineffective, yet things were bound to change over the subsequent decades.

---

[47] Cf., e.g., LaPierre (1934), Smith & Terry (2012) for a discussion of LaPierre's early study, and Kraus (1995) for an influential meta-analysis of research on attitude-behavior consistency. This issue will be discussed in section 3.1 below.

[48] Nisbett (2016: 67, my emphasis).

[49] Cf., e.g. von Neumann & Morgenstern (1944).

[50] Cf. Oliver (2017, Chapters. 1, 2, and 3) for an accessible introduction to the history and fundamental landmarks of this development within economic theory.

[51] Oliver (2017: 16).

A key figure in enabling this transition was the American polymath and future Nobel laurate in economics Herbert Simon, whose notion of *bounded rationality* was highly instrumental in raising economists' awareness towards the game-changing idea of bringing psychological insights and findings to bear more closely on the study of human economic behavior. In this regard, a central tenet of the standard model is that humans should and will maximize utility. In distancing himself from this increasingly shaky empirical assumption, Simon famously argued that, on the contrary, people by and large are not optimizers, but rather satisficers.[52] Differently from the ideal agent postulated by standard economic models, flash-and-bones humans – i.e., the kind of economic agents whose decisional patterns Simon was keen on better understanding – typically lack both the time and the ability to engage in the often complex calculations prescribed by rational choice theory, as well as the information needed to achieve utility maximization. As a consequence, he proposed, when faced with the myriad choices that they are required to make on a daily basis, people normally do not 'calculate', but rely instead on quick and dirty decision procedures know as *heuristics* – i.e. general rules of thumb or mental shortcuts that we (mostly automatically) apply in order to reach subjectively satisfactory and often roughly correct, albeit objectively sub-optimal decisions. This brilliant seminal idea soon proved a very powerful theoretical framework to accommodate the slowly accumulating evidence indicating that humans, as a matter of fact, abundantly violate the axioms of rational choice theory. Such violations, moreover, are not at all random, but systematic. This suggested that, contrary to what most economists kept assuming, real agents could no longer be plausibly described as behaving 'as if' they were following, albeit imperfectly, the rules of rational choice theory.[53] In other words, it was becoming increasingly clear that, when making their everyday decisions, people are not doing a poor job at following the prescriptions of rational choice theory, they are rather doing a remarkably good job at abiding by an entirely different set of rules – to wit, heuristics.

Starting from the 1960s, a group of psychologists started building on Simon's insights, and their work eventually converged with that of earlier unorthodox economists to form the theoretical and empirical bedrock of current behavioral economics. It is was in this new and stimulating ('formerly dismal') context that, in the early 1970s, Daniel Kahneman and Amos Tversky began to experimentally investigate and theorize about the many different heuristics that we rely on to navigate our environment, as well as to shed light on the many subtle ways in which these mental shortcuts can bias our judgments, such as, e.g., the status quo bias, representativeness, availability, anchoring, overconfidence, and confirmation bias, to mention just a few prominent ones.[54] For our purposes, however, a crucial aspect of Kahneman and Twersky's work is their demonstration of robust *framing effects* on human judgment. A basic assumption of standard economics is indeed that our preferences are fixed and stable. What Kahneman and Twersky repeatedly observed, on the contrary, is that, when faced with a choice among a given set of options, we display a marked tendency to change our preferences depending on the particular way in which those same options are *presented* or *described* to us – i.e. a violation of what is formally known as 'descriptive invariance'. This suggest that, contrary to classical economics, our preferences are not at all, as it were, etched in our minds once and for all, but rather constructed each time we make a choice. In other words,

[52] Cf. Simon (1955, 1956, 1957). For the notion of *bounded rationality*, see Wheeler (2018).

[53] Cf. Friedman, M., Savage, L. J. (1948).

[54] Cf., e.g., Kahneman et al. (1982), Ariely (2008), and Kahneman (2011).

they are heavily dependent on the overall structure of our decisional environment, or – as nudge theorists prefer to call it – *choice architecture*.

The idea of applying what we now call behavioral insights to policymaking is not a new nor even a recent one. Public policies inspired by psychological intuitions have been around for a long time. Limiting our considerations to the modern era, David Halpern provides a curious early example from 18th century Europe.[55] Many of us today appear to be in the grips of a virulent historicist variant of the so-called *rosy retrospection bias*,[56] which leads us to represent the 'good old days' as 'clearly' better than the present ones. The truth is, however, that the 1700s were brutally harsh times in central Europe for most human beings. The population was growing, and recurrent famines often led to bloody revolutions. Then came the potato, a recently introduced tuber that could potentially save millions from mass starvation. European leaders had just one problem: people would stubbornly refuse to grow and eat it (the thing was unfamiliar, tasteless, and the Bible made no mention of it). Many rulers passed largely ineffective laws aimed at imposing its adoption. One of them, Frederick the Great of Prussia, nicknamed "Old Fritz", initially did the same and, during a famine, doubled down by threatening to cut the noses and ears off any peasant who would not plant potatoes. Still nothing. Better to starve, apparently, than to be fed dog food! It was at this point that Old Fritz realized that a tad of psychology could do the trick. He first set the stage by publicly showing off a personal interest for the new food and its lovely flowers, then ordered his soldiers to display a heavy and visible guard around his (royal!) potato fields, but instructed them to be lax in protecting them. The rest is history (or, most likely, legend). Peasants started to sneak into the royal fields to steal potatoes, and the tuber soon began to be widely grown and eaten.[57] The moral, for us, is twofold. First of all, Halpern's historical anecdote is a vivid illustration of the now widely recognized fact that softer policy approaches to behavior change often pay off. Secondly, the example arguably lends further credibility to my picture of humans as natural born influencers. In particular, it suggests that – as Halpern himself puts it – "People have been 'nudging' each other for as long as mankind has existed. We are always busy persuading and encouraging those around us to do one thing or another."[58]

In spite of the powerful impact of human opinions and behavior on the final outcome of most public policies, however, psychological research only came to establish a systematic relationship with policymaking in the early 2000s. Starting from around this time, the idea that behavioral economics, and behavioral science more broadly, might be systematically relied upon in order to inform the design of public policy began to spread on a large scale.[59] A milestone in this process was the publication in 2008, by the economist Richard Thaler and the legal scholar Cass Sunstein, of the hugely influential book *Nudge: Improving Decisions about Health, Wealth and Happiness*.[60] Here is how Adam Oliver, a leading scholar of behavioral policy, describes some of the reasons that led to the book's immediate success:

---

[55] Cf. Halpern (2015:15-16).

[56] Cf. Mitchell et *al.* (1997).

[57] It is worth noting that this historical anecdote is a prime example of the efficacy of what Robert Cialdini, a leading scholar in the psychology of persuasion, refers to as the *scarcity principle* – i.e. the tendency of people to assign more value to goods that are less available or more difficult to obtain. Cf. Cialdini (2021, ch. 6).

[58] Halpern (2015:13).

[59] Cf., e.g., Camerer et *al.* (2003), and Thaler & Sunstein (2003).

[60] Thaler & Sunstein (2008, 2021).

> "In the same year that Thaler and Sunstein published their book, much of the world experienced the most serious financial crisis since the 1930s, which was caused substantively by insufficient regulation of the banks due to too much faith being placed in neoclassical economics and rational choice theory. Some began to look to behavioral economics for answers. Moreover, liberal-minded politicians, from the left and right of the political spectrum … were searching for ways in which they could motivate people to change their self and society-harming behaviours, without imposing further regulations or bans … Additionally … they were trying to identify policies that would be inexpensive to implement."[61]

According to this now standard narrative, in the wake of a catastrophic financial crisis that ultimately had its roots in a long series of reckless human choices, politicians were in dire need to explore new ways of changing people behaviors. They further wanted such ways to be non-mandatory, and cost effective. Feeling let down by standard economics, they started looking at behavioral economics for answers. One of those answers was nudging.

According to the distinction that I suggested above, nudges would count as external routes to behavior change, as they do not work by directly intervening on people's beliefs and attitudes, but rather by carefully reshaping their environments. The fundamental psychological insight that nudgers capitalize on, in particular, is that such environments must be adapted to the kind of decisional procedures that humans *actually* adopt – e.g. heuristics – as opposed to the ones that they *ideally* ought to adopt according to the standard model of rational choice. The basic idea of the nudge approach is indeed that by systematically taking into account our knowledge of humans psychological makeup, one can affect their behaviors by means of small adjustments in their so-called *choice architecture* – i.e. the context in which they make their decisions. Here is how Thaler and Sunstein themselves characterize a nudge:

> "A nudge … is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid."[62]

Nudges then basically work not by changing – i.e. adding or eliminating – the options that are presented to an agent, but rather by reorganizing the *same* set of options in a different way. These behavioral tools have so far been successfully applied across the world to many different policy domains – ranging from politics to finance, and, as we shall shortly see, the environment.[63] As of 2014, 136 out of 196 countries were already systematically relying on nudges in their public policies.[64] A standard example comes from the health domain, where nudging interventions have proved effective in altering people's self-harming dietary behaviors.[65]

---

[61] Oliver (2017: 108-9).

[62] Thaler & Sunstein (2021: 8).

[63] For examples of the use of nudges in public policy in the European Union and its member states, see Sousa Lourenço et *al.* (2016). For examples in the United States, see the *Social and Behavioral Sciences Team: Annual Report* (2016). For examples in the United Kingdom, see Halpern (2015). For examples in developing countries, see World Bank (2015).

[64] Whitehead et *al.* (2014).

[65] Cf. Rozin et *al.* (2011), Dayan & Bar-Hillel (2011), Arno & Thomas (2016), and Bucher et *al.* (2016).

Over the last few decades, policymakers, especially in the United States, have been increasingly concerned about the steady growth of obesity, a condition known to constitute a major leading cause of preventable deaths. There is hence an urgent need to reduce people's calories intake by encouraging the regular consumption of low-calorie food. Studies however show that the most common treatment for obesity – i.e. dieting – is, alas, largely ineffective. As a consequence, an alternative strategy currently endorsed by researchers consists in shifting the focus of attention from the individual herself to her environment – in our terms, from internal to external routes to behavior change. Very promising results, in this regard, come from interventions based on subtle modifications of cafeterias' overall layouts – i.e., specifically, on altering the way in which food is displayed in cafeteria lines. There are of course several possible ways of arranging the very same food options. French fries and carrot sticks, for instance, can be placed first, last, or in an altogether separate line, thereby making these food items more or less visually salient and accessible. Research shows that such small changes are far from inconsequential. Early work on obesity indeed suggested that manipulating the effort required to reach food often has surprising effects.[66] In line with this prediction, recent experiences demonstrate that simply rearranging a cafeteria line by subtly, and often unnoticeably, varying the locations of several food items – without in any way altering people's choice set – has significant effects on the kind of food choices that people end up making. Rozin and his colleagues (2011), for instance, found that making a given food item less reachable by moving it of about 25 cm. (10 inches) or making it harder to transfer to one's own plate by changing the serving utensil (tongs or spoon) can reduce its consumption by 8 – 16%.[67]

The above example gives us the opportunity to highlight two crucial aspects of nudges as policy instruments. To begin with, nudges are usually conceived as *libertarian* or liberty-preserving policy instruments, as they are intentionally designed to at least ideally preserve – and indeed, make easier to exercise – the individual's freedom of choice. According to nudge theorists, people should always be let free to go their own way without incurring excessive burdens to do so. With respect to the above example, for instance, it seems that making French fries slightly less visible or easy to reach clearly won't prevent anyone from helping herself to a serving of this tempting food.[68] At the same time, however, nudges constitute a *paternalistic* policy instrument, as reliance on them is premised on the idea that people often make demonstrably poor decisions. Thaler and Sunstein like to express this idea by comparing the imaginary species *Homo economicus* (or Econs) – that populates standard economics textbooks – to the real species targeted by choice architects: *Homo sapiens*, or Humans. Contrary to Econs, they are keen on reminding us, our ilk is heavily biased, and typically cannot "think like Albert Einstein, store as much memory as Google … or exercise the willpower or Mahatma Gandhi."[69] It follows, in their view, that it is morally legitimate for – and often, indeed, even morally required of – the policymaker to deliberately

[66] Stanley Schachter's theory of obesity, for instance, predicted that obese individuals would reduce food intake relative to normal weight individuals as a function of the effort required to access food, and this prediction has been confirmed by many subsequent studies conducted both inside and outside of Schachter's framework. Cf. Schachter (1971).

[67] Cf. Rozin et *al.* (2011).

[68] Cf. the following, much quoted, passage from *Nudge*: "Nudges are not taxes, fines, subsidies, bans or mandates. Putting the fruit at eye level counts as a nudge. Banning junk food does not" (Thaler & Sunstein 2021: 8).

[69] (Thaler & Sunstein 2021: 9).

influence people's behavior, in order to make their lives healthier, wealthier, and happier.[70] This position, known as *Libertarian Paternalism*, is however intended as a soft and nonintrusive kind of paternalism – and typically characterized as a paternalism of means, not of ends – as the nudgee (the person being nudged) is seen as the ultimate judge of any given intervention's legitimacy. Nudges, so the slogan goes, should make choosers better off *as judged by the choosers themselves*.[71] Ideally, then, nudges constitute a very profitable tool that can help individuals make choices that they themselves, on reflection, would consider better. As we shall see in the next chapter, not everyone shares the same enthusiasm with respect to nudges potential as gentle instruments of human enhancement. Before we turn to ethical considerations, however, let us first take a look at what nudges can do to help solve problems in a policy domain that we should all be increasingly concerned with: the environment.

1.5. Better is good

Human attitudes and behaviors are at the heart of many increasingly worrying environmental problems, the most pressing among which is currently climate change. In spite of what the usual 'merchants of doubt'[72] would have us think, there exist today not only a long standing and broad scientific consensus on the reality of anthropogenic climate change, but also a more recent scientific consensus on the very existence of the latter.[73] Global warming is happening fast, it is directly linked to our individual and collective choices, and it requires urgent action. It arguably follows that listening to what behavioral scientists have to say on the matter, and letting our environmental policies be informed by their best findings and insights, is definitely a good idea. Keeping considerations at the individual level, a preliminary distinction can be drawn among pro-environmental behaviors aimed at the mitigation of or adaptation to the current climate crisis, according to whether their impact on climate change is either direct or indirect. To exemplify: producing, gathering or disseminating reliable scientific data on global warming – just as funding, organizing or attending climate summits – are all behaviors ultimately aimed at combating climate change, though only *indirectly* so. The efforts of behavioral scientists, however, have so far mostly focused on behaviors that *directly* result in a measurable reduction of emissions, such as, e.g., reducing water or energy consumption, recycling, keeping air travel at a minimum, or driving fuel efficient cars. I will accordingly focus on the latter, and collectively refer to them as *climate-change-responsive behaviors* (CRBs).

As we noted in the last section, people's beliefs and attitudes about various issues are not at all easy to sway one way or another. Pro-environmental attitudes – usually defined and understood in terms of concern for the environment or caring for environmental issues – appear to be no exception in this regard, as they are usually deeply intertwined with our own lifestyles, values, and overall worldviews. Indeed, although these constructs are now known to be affected by a wide array of factors – the most investigated amongst which currently being demographics (such as,

[70] With respect to food choices, for instance, the author of *Nudge* observe the following: "We do not claim that everyone who is overweight is necessarily failing to act rationally, but we do reject the proposition that all or almost all people are choosing their diet optimally" (Thaler & Sunstein 2021:9-10).

[71] (Thaler & Sunstein 2021:9-10).

[72] Oreskes & Conway (2010).

[73] Cook et *al.* (2016).

e.g., age, gender, socio-economic status, and country), religion, politics, values, direct experience, and knowledge – researchers at present hold widely different views concerning both their internal structure, and (partly as a consequence of this fact) the most reliable ways to measure them.[74] Moreover, as we also noted in section 1.4 above, a long tradition of research on attitude-behavior consistency has amply shown that, while attitudes are indeed often predictive of behavior, there remain important exceptions to this general rule – i.e. explicitly held attitudes do not always lead to congruent behaviors and actions. This seems especially true in the case of CRBs, a domain of human activity which – just as, e.g., safety behavior (wearing helmets, safety belts) or health behavior (diet, smoking) – has proved strongly resistant to behavior change. Whereas strong links are indeed typically observed to hold between pro-environmental attitudes and *self-reported* (usually, in fact, overreported due to a *social desirability* bias[75]) CRBs, much weaker links have typically been found between the former and *observed* CRBs. This phenomenon may well be partly due to the fact that explicitly held pro-environmental attitudes, being usually measured by means of self-report scales, are bound to be powerfully affected by a strong social desirability bias.[76] At the same time, however, this is unlikely to be the whole story about such a complex issue.

A long-investigated phenomenon, in this regard, is the so-called *Value-Action Gap* – i.e. the large, systematic, and worrying disconnect between our possession of environmental knowledge and awareness, on the one hand, and our adoption of CRBs, on the other.[77] As we will see in the next chapter, for instance, the vast majority of us today know – and, if (anonymously!) queried on the matter, resolutely assert – that saving on household energy is currently in the best interest of both present and future generations, yet only a very limited number of people who can afford it actually end up choosing to install solar panels or to rely more heavily on renewable energy sources. The same holds for countless other CRBs. "Simply put" – as it has recently been observed – "there is an astounding difference between environmental attitudes people harbor and environmental behaviors we engage in."[78] Why exactly is this the case? What prevents us from 'doing the right thing' both in the long term, and on a daily basis? Environmental psychologists, behavioral economists, and other social scientists have long been wrestling with this elusive question in the hope of locating the roots of the problem. In doing so, they have so far isolated a host of factors that act like barriers to the widespread adoption of CRBs. Some of these barriers are normally seen and treated as *structural* ones – i.e. to fall beyond the individual's direct control. Low income, for instance, can prevent me from installing solar panels, and living in a rural area where public transport is sparsely available can leave me no options but to drive my kids to school, and myself to work. Structural barriers must arguably be removed by means of legislation and careful urban planning. Other barriers, however, are not structural but *psychological* in nature – i.e. regarded as falling, at least in theory, within the boundaries of the individual's direct control. What are they,

---

[74] Gifford & Sussman (2012).

[75] Cf., e.g., Bord et *al.* (1998). But see also Milfont (2008) for results that vindicate the reliability of self-reported measures in the face of social desirability.

[76] Gifford & Sussman (2012).

[77] Numerous field studies have documented this phenomenon. Cf., e.g., Fahy (2005) for the case of waste management, and Flynn et al. (2009) for the case of sustainable energy use. Kollmuss & Agyeman (2002) review the most common and influential analytic frameworks that have so far been used to explain the gap, as well as the demographic, external, and internal factors that have been found to influence CRBs. Pirni (2023) analyzes the motivational gap with a particular focus on climate change, and indicates a twofold path to overcome it. Cf. also Blake (1999), and Croteau (2019).

[78] Croteau (2019: 21).

and, most importantly, what can be done to remove them? With respect to the first question, the most comprehensive and fine-grained analysis to date is perhaps to be found in the decades-long work of Robert Gifford, a leading environmental psychologist.[79] Gifford (2011) singled out 29 crucial factors that have been shown to increase our apparent amotivation to act, and that he therefore evocatively calls the 29 'dragons of inaction'. These factors have then been grouped by the author into seven main categories, interacting with each other to bring about what Thaler and Sunstein have dubbed the 'perfect psychological storm'.[80] Let us then take a selective look at some of Gifford's dragons.

An interesting factor in Gifford's list falls under his category of 'limited cognition'. As we saw above, a general assumption of behavioral economics is that Humans, as opposed to Econs, are not optimizers, but satisficers. One way in which this crucial aspect of our psychological makeup can interfere with our consciously recognized environmental goals is by affecting our responses to both *uncertainty* itself, and messages appealing to this concept. Research on resource dilemmas indeed shows that real or perceived uncertainty makes the adoption of CRBs less likely.[81] As a consequence, it is reasonable to expect that *perceived* uncertainty about climate change will itself function as a justification for inaction. An interesting finding, in this regard, is the framing effect observed in people's responses to the typical level-of-confidence talk (e.g. "likely", "very likely") contained in the 2007 IPCC report. This language apparently led many individuals to interpret the climate change scenarios described in the report as being less likely than its authors intended.[82] In a way, this is hardly surprising. As I put it at the beginning of section 1.3, humans are hopeless control freaks – i.e. we like or need to bask in the reassuring thought that our lives and fortunes are a lot more under our control that it is actually the case. Perhaps, to a certain extent, this is a good thing, as it is arguably part of what gets us out of bed in the morning. Yet one of its many epistemically unwelcome consequences, I take it, is that makes us all suckers for certainty – i.e. we *want* to be told that, if we act in specified ways, things will *surely* go one way or another. The rub, however, is that, due to their training, scientists of all stripes will forever frustrate our expectation by carefully shunning all certainty talk. Their common stock-in-trade, after all, are not absolute truths, but rather confidence intervals. Scientific knowledge, indeed, is typically model-based, and climate scientists, just as other scientists, can only rely on their best models in order to make predictions that would only inappropriately be described as simply true or false, as opposed to more or less likely to be (partly) confirmed. The general upshot of this human state of affairs is that climate scientists' epistemically praiseworthy attempts to fairly characterize the certainty degree warranted by their data may in effect lead to a dangerous underestimation of climate-related risks on the side of their lay audience. Climate scientists, then, seem to face a nasty ethical dilemma: scare the heck out of people in the interest of our planet by going well beyond what their data allows them to say, or present the data objectively, thereby running the risk of encouraging unsupported, and highly consequential optimism? It may well be due to my cultural

[79] Cf. Gifford (2011, 2014), Gifford et *al.* (2009), Gifford et *al.* (2011), Gifford & Nilsson (2014), Lacroix & Gifford (2017). Some of the barriers identified by Gifford and colleagues are also discussed in Kollmuss & Agyeman (2002), Lorenzoni et *al.* (2007), and Thaler & Sunstein (2021, ch. 14).

[80] "Sadly, we have what amounts to a perfect storm, a confluence of factors that make collective action difficult." (Thaler & Sunstein 2021: 282). The factors that the authors have in mind are of course psychological or better behavioral in nature.

[81] Cf., e.g., Hine & Gifford (1996).

[82] Budescu et *al.* (2009).

upbringing, but I would personally go for the second option, coupled with massive efforts aimed at increasing the currently insufficient level of scientific literacy.

Another interesting psychological barrier highlighted by Gifford is represented by peoples' beliefs about how much control they have over the outcomes of their actions (*perceived behavioral control*), and about the impact that their actions have on the environment (*perceived self-efficacy*). In section 1.2, while considering some of the main findings of persuasion research, I noted that, in the case of health campaigns, the effectiveness of messages appealing to fear appears to be a function of a subject's perceived self-efficacy – i.e. persuasion attempts appealing to this powerful primal emotion (fear) seem to be only effective to the extent that the subject is also provided with specific instructions about what to do in order to avoid the undesirable consequences of their behavior. The subject's underlying reasoning, I suggested, might be driven by the stoic thought that there is no point in worrying about something when there is little one can do to avoid it. Now, the same line of reasoning seems to be even more powerfully activated when confronted not with an individual, but rather with a global problem such as climate change. Apparently, we sometimes fail to adopt CRBs precisely because we experience a lack of control over the outcome of our actions. Heath and Gifford (2002), for instance, relied on an extended version of Icek Ajzen's *theory of planned behavior* (TPB)[83] in order to predict and explain the public transportation use patterns among students of a medium-sized university in western Canada. One of the things they found is that an experimental manipulation of students' perceived behavioral control may not just lead to a significant increase in public transport use, but also positively affect students' beliefs and attitudes concerning transportation options (this latter finding that will prove relevant in chapter 4).[84] Illusory or not, then, our sense of being in control of our actions is apparently a strong predictor of our propensity to engage in a crucial CRB. Interestingly, current research further suggests that the message-framing issue broached in the last paragraph is also relevant to the managing of many peoples' perceived lack of self-efficacy. In a telephone survey ran among Ontario residents, Gifford and Comeau (2011) found that *motivational* messages (couched in terms of the relational "We" – e.g. "my neighborhood will be a healthier place to live if *we* walk more to cut greenhouse gasses") as opposed to *sacrifice* messages (couched in terms of the individualistic "I" – e.g. "*I* am going to have less freedom to make the choices *I* want if we are going to solve climate change") positively affected subjects' perceived climate change competence, engagement, and behavioral intentions.[85] Empowering (motivational or value-oriented) messages, then, can apparently attenuate peoples' perceived lack of self-efficacy.

A final barrier well worth noting for our purposes is the one induced by the hard to exaggerate power of *social comparison* in our lives. In spite of our occasional, and slightly pathetical, smug remarks to the contrary, we do care a great deal about what other people are thinking and doing. As of 2022, the largest social networking service on this planet, Facebook, claimed to have nearly three billion active users.[86] Indeed, as (honestly-reported) personal experience suggests, and social psychologists have long acknowledged, comparing our actions to those of others is one of the

[83] Cf, e.g., Ajzen (1991). We will go back to this approach in section 3.1 below.

[84] Heath & Gifford (2002). Similar findings were reported by Kaiser & Gutscher (2003), also working within the framework of Ajzen's TPB.

[85] Gifford & Comeau (2011).

[86] Meta Reports Fourth Quarter and Full Year 2022 Results. Accessed on November 2024 at: https://investor.fb.com/investor-news/press-release-details/2023/Meta-Reports-Fourth-Quarter-and-Full-Year-2022-Results/default.aspx

most powerful drives in human psychology. The first two predictions made by Leon Festinger's seminal *theory of social comparison processes*, for instance, read respectively as follows: "*Hypothesis I*: There exists, in the human organism, a drive to evaluate his opinions and his abilities"; "*Hypothesis II*: To the extent that objective, non-social means are not available, people evaluate their opinions and abilities *by comparison* respectively with the opinions and abilities of others."[87] Importantly, from an early age, social comparisons are often carried out in order to draw from the observation of our peers substantive *normative* information about what such peers regard as the right thing to do under various circumstances. This aspect of human psychology, as we shall presently see, has been systematically exploited by nudgers to encourage the adoption of CRBs. Within the present context, however, it must be noted that our natural tendency to alter our behavior in order to fulfil the normative expectations of our social group makes norms a double-edged sword, as their effects on our behavior will inevitably reflect the attitudes and behaviors that happen to prevail in our reference group. In a recent book, economist Robert Frank has extensively explored the effects and policy implications of what he calls *behavioral contagion* in various domains of human activity.[88] One of these domains concerns energy-related behaviors – in particular, the decision to invest on solar panels. This choice appears to be strongly influenced by mechanisms of peer comparison. Graziano and Gillingham (2015), for instance, have studied the spatial patterns characterizing the distribution of solar panel installations across the state of Connecticut, U.S. Factoring out variables such as, e.g., household income, they found a significant *neighbor effect* on residents' decision to install solar panels. Now, it may well be the case that at least some of the people who finally resolve to cut back on their grey energy use, do so after long hours of soul-searching and careful deliberation. There is also evidence, however, that whether and when your neighbor decided to go for solar energy is a strong predictor of whether and when you will decide to follow suit. Insofar as that is the case, living in a neighborhood where environmental concern is not the norm, may constitute an invisible barrier to behavior change.

For the sake of brevity, the one above was just a small sample of the many psychological barriers that have been found to stand in the way of our decisions to engage in CRBs. As anticipated, the next question on our agenda is what can be done to remove them. In this regard, as policymakers across the globe have been increasingly acknowledging, nudging represent today a remarkably effective, and highly cost-effective, external route to behavior change. As a consequence, most governments and institutions now regard so-called *green nudges* as a key tool in combating climate change by promoting the adoption of various CRBs.[89] In particular, as we shall see, nudges appear to constitute a promising policy instrument for bridging the already discussed Value-Action Gap by helping people to realign their environmental decisions with what they already appear to regard as the right thing to do. Thaler and Sunstein, as we saw in section 1.4, think that nudges should help individuals make choices that they themselves, on reflection, would consider *better*. This however raises an obvious question – i.e. better *for whom*? – that allows us to introduce an often-drawn distinction in the nudging literature. Green nudges are indeed usually considered *pro-social* – as opposed to *pro-self* – nudges, as they are not primarily aimed at improving the welfare of the nudgee (i.e. the person being nudged), but rather at reducing negative environmental externalities,

[87] Festinger (1954: 117, 118, emphasis added).

[88] Frank (2020).

[89] On green nudges as policy instruments, see, Beckenbach & Kahlenborn (2016), Elberg Nielsen et *al.* (2016), Schubert (2017), Evans et *al.* (2017), Wynes et *al.* (2018), Carlsson et *al.* (2019), Croteau (2019), Behavioral Insight Team (2020), Moratti (2020), Bonini & Dorigoni (2024) and Singh et *al.* (2024).

such as, e.g., air pollution or water waste.[90] Although the number of green nudges that have so far been implemented is growing by the day, a useful taxonomy has been proposed by Christian Schubert.[91] Schubert (2017) distinguished three paradigmatic types of green nudges, according to the specific psychological mechanism each one of them exploits. I will accordingly refer to them as Type 1, 2, and 3. The three mechanisms that most green nudges capitalize on, according to Schubert, are our desire to maintain a positive self-image (Type 1), our inclination to 'follow the herd' (Type 2), and our tendency to stick with preset default options (Type 3). As this latter family of green nudges will be discussed in chapter 2, I will conclude the present chapter by briefly considering the first two families of green nudges (henceforth simply 'nudges') singled out by Schubert's taxonomy, and illustrating them by means of examples falling into each.

Type 1 nudges, as we just saw, aim at encouraging the adoption of CRBs by leveraging our natural desire to feel good about our decisions and actions. Everyday consumer choices, as it has been observed, are indeed a way in which modern society allows us to express the values that we take to define our lives, and to cultivate a positive self-image by acting according to those values.[92] As Schubert points out, there are at least three ways in which choice architects can and have helped us in achieving this goal. The first one basically consists in making it easier – i.e. facilitating the adoption of CRBs by reducing the *cognitive costs* typically associated with many consumer choices. This can be done, for instance, by simplifying the often-complex way in which information about various environmentally-relevant product features is presented. The guiding idea here is that such information should be structured or framed in order to better match Humans' – as opposed to Econs' – processing abilities and limitations.[93] A second, and related strategy consists in attempting to raise the level of consumers' *environmental awareness* by increasing the *salience* of those same product features – i.e. making them stand out with respect to other, environmentally-irrelevant ones. This can be achieved thorough the implementation of so-called *eco-labelling*. Eco-labelling makes for a Human-friendly kind of feedback provision in that – unlike the standard, and Econs-tailored product labelling – it does not capitalize on the empirically flawed assumption that real consumers always pay sufficient attention to the long-term consequences of their choices. Although this strategy can potentially be applied to all sorts of consumer goods, a telling example involves cars, where eco-labels can be exploited to make cars' fuel-efficiency more prominent.[94] As it is the case with other types of nudges, however, it is important to note that eco-labelling can at times backfire. A case in point here is the mandatory energy labelling scheme for electrical appliances, introduced by the EU in 1995. The scheme rated appliances' energy efficiency on a seven-point colored scale (i.e. A-green to G-red). By 2003, however, about 90% of appliances had reached level A, which led policymakers to redesign the scheme by introducing finer-grained levels within the former A (i.e. A+ to A+++). In hindsight, this move has proved ill-considered. Ölander and Thøgersen (2014) indeed found that, at least with respect to TV sets, the revision actually made customers *less* likely to buy the most energy-efficient appliances – most likely due to the fact that they had started to take A as the only relevant reference point, and to

[90] Economists define an *externality*, in general, as "a cost or benefit incurred or received by third parties who have no control over its creation" (Frank 2020: 27). For an informed discussion of the distinction between *pro-self* and *pro-social* nudges, see Congiu & Moscati (2021).

[91] Schubert (2017).

[92] Cf., e.g., Sunstein & Reisch (2014).

[93] Cf., e.g., Lehner at *al.* (2016).

[94] Cf., e.g., Sunstein (2014, ch. 2).

therefore regard all the new sub-categories as more or less identical.[95] A final way to help people maintain a positive self-image by adopting CRBs consists in harnessing their sense of *social identity*. One of the most successful examples, in this case, is the 'Don't mess with Texas' social advertising campaign, designed to reduce the amount of littering along Texas's highways. Appealing to people's sense of community pride (a 'true' Texan would never do that!) clearly pays off. The now world-famous campaign is indeed estimated to have reduced littering by about 70% between 1986 and 1990[96], and policymakers are now routinely leveraging what is often referred to as *identity-based cognition* in order to foster various kinds of socially-beneficial behaviors, such as, e.g., using public toilets in India or wearing face masks during the COVID epidemic in Montana.[97] Evidently, as Robert Cialdini, a leading scholar of persuasion, would put it: "the "We" is the shared me".[98]

Let us then move on to Type 2 nudges – i.e. the ones that aim at fostering the adoption of CRBs by capitalizing on our inclination to 'follow the crowd' or imitate the behavior of our peers. As we noted above, comparing our beliefs and actions to those of others is one of the most powerful drives in human psychology, a fact well known to social psychologists at least since Solomon Asch's famous conformity experiments.[99] Moreover, evidence today suggest that the power of social influences on our CRBs remains largely undetected by the people affected by them.[100] It is hence far from surprising to discover that Thaler and Sunstein regard social influence a "one of the most effective ways to nudge (for good or evil)".[101] In this regard, a nudging strategy that has proved remarkably effective in steering people towards wiser environmental behaviors consists in leveraging their desire to conform to social expectations by conveying *social norms* in various ways. The norms in question can be either explicitly *prescriptive* – i.e. specify which choices are approved or disapproved within a relevant reference group – or, more often, merely *descriptive* – i.e. simply inform us that a particular choice happens, for whatever reason, to be prevalent within that group. Two famous green nudges that adopt this strategy target energy conservation and indirect water conservation respectively. The first one involves Opower, a U.S. based company that in 2008 started regularly mailing to their customers home energy reports (HERs). HERs contain information about the household's energy use, and, crucially, about how this use *compares* to that of their neighbors with similar-sized homes and heat type. Allcott (2011) reports that the *peer comparison* effect triggered by the nudge led to a 2% reduction of energy consumption – and effect that he estimated to be equivalent to that of a short-run electricity price increase of 11 to 20%.[102] The second example was implemented in two field experiments designed to assess how social norms can indirectly affect water conservation by avoiding unnecessary laundering. The nudgers goal, in this case, was to induce hotel guests to reuse their towels by exposing them to signs, placed in bathrooms, appealing to descriptive norms. Goldstein and colleagues (2008) found that appeals to descriptive norms (e.g. "almost 75% of guests reuse their towels"), as

[95] Ölander and Thøgersen (2014).
[96] Mols et *al.* (2015).
[97] Cf., e.g., Thaler & Sunstein (2021: 79-81).
[98] Cialdini (2021).
[99] Cf., e.g., Asch (1951, 1956).
[100] Cf., e.g., Nolan et *al.* (2008).
[101] Thaler & Sunstein (2021: 65). And perhaps even less surprising to find out that the latter author has dedicated a book to this theme. Cf. Sunstein (2019).
[102] Cf. Allcott (2011). Cf. also Costa & Kahn (2013), and Allcott & Rogers (2014).

opposed to, currently standard, general appeals to environmental protection (e.g. "help save the environment by reusing your towels") led to a 9% increase in the reuse rate.[103]

As we already noted, social influence, by its very nature, is a double-edged sword, and the effects of peer comparison on our adoption of CRBs are largely a function of the kind of behaviors that happen to prevail in our social group. It is hence, again, unsurprising to discover that both of the nudges just considered – just as countless others – can at time backfire and trigger unwelcome boomerang effects. However, as these and other psychological aspects of the nudging process will be discussed in chapter 3, let for now simply reiterate that, as the above examples stand to show, green nudging represents today a remarkably effective and highly cost-effective tool of environmental policymaking. Moreover, as it is the case in other policy domains, and as typically stressed by their supporters, green nudges are not (and should not be) intended as replacing or competing with, but rather as usefully complementing more traditional regulatory measures, such as incentives, bans and mandates. In particular, it is obviously unrealistic to expect that green nudges will constitute the definitive solution to the massive and extremely complex problem of climate change. This being said, it would arguably be deeply wrong-headed and empirically unsupported to therefore suggest that choice architects have nothing valuable to contribute to this increasingly pressing, and increasingly worrying, human cause. It hence seems appropriate to conclude this chapter by quoting Thaler and Sunstein on this matter:

> "Now, … if you are hoping to hear that you shouldn't worry because low-cost nudges will make this problem go away, we are going to disappoint you. As we have often stressed, not all problems can be solved with light-touch interventions … but there are also a wide variety of useful interventions that fully qualify as nudges that should be part of the environmental tool kit. By themselves, those steps will not eliminate the risks of climate change. But they will help, and as former president Obama likes to say about initiatives that merely dent large-scale problems: "Better is good."[104]

[103] Cf. Goldstein et al. (2008). Yet see Bohner & Schlüter (2014) for a replication failure of some of the findings reported by Goldstein and colleagues. We will go back to this issue in section 3.3 below.

[104] Thaler & Sunstein (2021: 286, 301).

2. Nudges and Rational Agency

2.1. To nudge or not to nudge

As we already saw in the first chapter, policymakers across the world currently regard nudges as a promising external route to behavior change. This route is normally not meant as substituting or as competing with, but rather as complementing other, more traditional policy instruments. The fundamental idea of nudge-based policies is that the quality of people's decisions can be demonstrably improved simply by changing the way in which the same options are presented to them – i.e. without outright banning any of those options or significantly altering their economic incentives. Nudges, that is, hold a liberal promise to steer people in predictable directions, while at the same time preserving their freedom of choice. This basic insight, as we have also seen in the last chapter, can validly contribute to tackle many increasingly worrying environmental threats that currently beset our beautiful planet. As a consequence, most governments and institutions, over the last couple of decades, have embraced the idea that green nudges happen to represent a very profitable tool in combating climate change – both on the mitigation and adaptation front – by encouraging the adoption of what I called climate-change-responsive behaviors (CRBs). Nudges, in other words, would seem to constitute a remarkably effective strategy to partly bridge the Value-Action Gap by helping people to realign their environmentally-relevant decisions and choices with what they already appear to regard as the right thing to do.

As their supporters often point out, at least three pragmatic considerations seem to clearly count in favor of the widespread adoption of nudge-based policies. For one thing, compared with most other policy instruments, nudges are highly cost-effective, as behavioral interventions can usually be implemented at virtually zero cost.[1] Another considerable advantage is that, contrary to other often unavoidable regulatory measures, such as taxation or fines, some evidence today suggests that – at least to the extent that choice architects are perceived as promoting goals that they themselves reflectively endorse – most citizens do not seem to mind being nudged.[2] A final pragmatic consideration typically put forward by nudge supporters is that, insofar as there is no neutral way of presenting options, "objecting to nudges per se" – in Thaler and Sunstein's own words – "makes as much sense as objecting to air and water"[3], as choice architecture is simply inevitable. In following passage, the second author brilliantly articulates this basic idea by further elaborating on the water metaphor:

> "Consider … a tale from the novelist David Foster Wallace: "There are these two young fish swimming along and they happen to meet an older fish swimming the other way, who nods at them and says "Morning, boys. How's the water?" And the two young fish swim on *for a* bit, and then eventually one of them looks over at the other and goes "What the hell is water?"" This is a tale about choice architecture. Such architecture is inevitable, whether or not we see it. It is the equivalent of water. Weather is itself a form of choice

[1] Cf., e.g., the host of successful policies presented in Halpern (2019).
[2] Cf., e.g., Sunstein (2016a: ch. 6).
[3] Thaler & Sunstein (2021: 313).

> architecture, because it influences what people decide. Needless to say, human life is not imaginable without some kind of weather. Nature nudges … In this light, *choice architecture is inevitable*. *Human beings* (or dogs or cats or horses) *cannot wish it away*."[4]

It follows, according to Libertarian Paternalism, that outright preventing governments from nudging people toward choices that are at least bona fide supposed to make their lives healthier, wealthier, and happier, would in effect be tantamount to letting private companies have a field day steering them toward choices that, while certainly increasing their profits, will instead make them less healthy, wealthy, and happy.[5]

To put it mildly, however, not everyone shares the same level of enthusiasm about the prospects of relying on nudges in public policy. While constituting a very promising tool to encourage the adoptions of CRBs, nudges have nonetheless proved rather controversial, and their use in policy has been put under close ethical scrutiny.[6] Many indeed worry, on various grounds, that nudge-based policies may end up clashing with widely held human values, such as liberty, autonomy, respect, and dignity. In this regard, most ethical concerns to date seem to focus on *autonomy*.[7] Different understandings of this highly multifaceted notion have so far emerged within the animated ethical debate on nudging, with different authors accordingly directing their intellectual labors to different aspects thereof. In their recent overview, Schmidt and Engelen (2020) have conveniently distinguished the following four main dimensions along which autonomy has been considered:[8]

- Freedom of choice
- Absence of domination
- Volitional autonomy
- Rational agency

Although the present chapter will be concerned with this latter understanding of autonomy, let me conclude the present section by briefly considering each one of the other three in turn. With respect to the first dimension, some authors have maintained that, contrary to Thaler and Sunstein's definition[9], nudge-based policies will inevitably end up undermining our freedom of choice – nudges, in their view, are not nearly as 'liberty-preserving' and 'easily resistible' as their supporters would have us think. Indeed, while some are rather skeptical of the idea that such policies can actually live up to their widely advertised libertarian credentials[10], some others have forcefully argued that Libertarian Paternalism is flat out incompatible with liberal principles. The driving idea here is that, even though supporters of this currently fashionable political doctrine have skillfully lured many people into thinking otherwise, their 'libertarian' brand of Paternalism,

---

[4] Sunstein (2015: 420-21, emphases added).
[5] Cf., however, Hausman & Welch 2010, and Vallier (2016) for criticisms of the inevitability argument.
[6] Cf., e.g., Sunstein (2016a), Schmidt and Engelen (2020), and Kuyer and Gordijn (2023).
[7] Cf., e.g., Bovens (2009), Hausman & Welch (2010), Grüne-Yanoff (2012), Saghai (2013), MacKay & Robinson (2016), Viale (2022), and Calboli & Fano (2022). Cf., in particular, Vugts et *al.* (2020) for a systematic literature review.
[8] Schmidt & Engelen (2020: 4). For expository reasons, the options have not been listed in the same order.
[9] Cf. Thaler & Sunstein 2021: 8.
[10] Cf., e.g., Rebonato (2014).

in spite of its name, would be just as manipulative, and it would in effect, upon closer scrutiny, boil down to good old Paternalism, disguised under a new dress. In particular, just as it was the case with its former instantiation, the new Paternalism 2.0. would also be culpable of implicitly relying on a notion of welfare that fails to take into due account, and hence to respect, the subjectivity and plurality of people's values.[11] As to the understanding of autonomy as 'absence of domination', there is a concern that, as a policy instrument, nudging may, as it were, fall in the wrong hands, where it could be then used to exercise an invisible, and hence especially dangerous control over our lives. In this regard, some have appealed to the explanatory framework provided by Foucauldian theories of power in order to argue that some modern governments are already relying on choice architecture to 'discipline' their citizens through their own seemingly free choices. Some others have instead appealed to neo-republican theories of freedom in order to argue that, due to its frequent lack of transparency, nudging runs the risk of eluding individual and democratic control, thereby making it too hard or even impossible to hold public officials fully accountable for the consequences of their actions.[12] Moving on to the third understanding of autonomy, some critics taking part in the debate have focused on the volitional – as opposed to the more strictly cognitive – component of this notion. The guiding principle, in this case, is that our decisions and choices, in order to be regarded as autonomous, should reflect our true preferences. Frankfurt (1971), for instance, famously argued that our ability to form *second-order desires* – i.e. desires about desires – is what ultimately and essentially distinguishes our own species from all other animals. "Besides wanting and choosing and being moved to do this or that" – as he put it – "men may also want to have (or not to have) certain desires and motives. They are capable of wanting to be different, in their preferences and purposes, from what they are."[13] Some hold, in this regard, that, contrary to choice architects stated intentions, nudges would not at all help people realign their decisions with what they on reflection regard as the right thing to do – i.e., arguably, their 'true' preferences or desires – but rather shut them off from their 'true' selves, and hence ultimately play a worryingly alienating function.[14]

2.2. The autonomy objection

As I see it, the most pressing objection to nudge-based policies is the one that sees nudges as a serious threat to our rational agency. Nudging, in the eyes of many, would indeed be bound to hinder or even positively undermine our ability to make autonomous decisions, and consequently act on them. The thought, in particular, is that, by relying on nudges, policymakers would fail to respect us qua rational agents by deceitfully capitalizing on our cognitive foibles – i.e. the host of heuristics and biases that have long been known to characterize human cognition. As choice architects, on this view, typically work their magic by leveraging our less-than-rational or even downright irrational decisional processes, their cunning brain hacks would fall short of counting

[11] Cf., e.g., Grüne-Yanoff (2012).
[12] Cf., e.g., Jones et *al.* (2011), and Pettit (2014) respectively.
[13] Frankfurt (1971: 7).
[14] Cf., e.g., Hausman & Welch 2010.

as morally legitimate policy interventions.[15] I will henceforth refer to this line of criticism as the *autonomy objection* (henceforth, AO).

AO has unsurprisingly fueled a heated debate, and different response strategies have so far been attempted on nudges' behalf. Some have defended their use on largely pragmatic grounds, either by reiterating the many (individually and socially) unwelcome consequences of human fallibility, or by stressing the unavoidability of choice architecture.[16] Others have endorsed a sort of *faute de mieux* defense – i.e., one that would justify, and therefore limit, the use of nudges only to cases where other, purportedly more agency-respecting policies – such as, e.g., education or moral suasion – prove ineffective.[17] A yet bolder move, to which I am largely sympathetic, has instead consisted in questioning the very notion of rationality that underlies AO. The latter has indeed been contrasted with a different, and arguably more realistic, view of human cognition, on which the host of automatic psychological mechanisms that choice architects often rely on, while perhaps not epistemically, would nonetheless still count as ecologically rational.[18] A final batch of authors have adopted what can be dubbed a 'mischaracterization' strategy, by pointing out that critics wielding AO are in fact relying on a rather skewed or at least incomplete understanding of nudges' cognitive workings.[19] The following considerations are meant as a contribution to this last response strategy.

As in the case of many other objections to the use of nudges as policy instruments, I take AO to be motivated by widely sharable concerns, and to therefore deserve a fair hearing. I do agree, in particular, that if nudges, upon closer scrutiny, turned out to represent a serious threat to our rational agency, then this fact alone, regardless of their effectiveness, would already call for an urgent reassessment of their widespread use in policy. I do believe, however, that AO can be shown to lose much of its initially intuitive appeal, and hence of its force, in light of the following considerations.

AO's gist, as I put it above, is that nudges, or better nudgers, would not be paying due respect to our rational agency, as their intended goal is to steer our choices by cleverly exploiting a plethora of allegedly irrational decisional processes. It seems to follow that any serious attempt at fleshing out this basic charge in more detail will call for an indication of where exactly the irrationality in question ought to be looked for. To clarify, let us agree on understanding a 'nudged choice' as the output of an often automatically-triggered decisional process taking place within the context of a choice architecture that has intentionally been designed in order to promote a given behavior. By my lights, the relevant question is then follows: what exactly is it about such a choice that would make it an instance of irrational (or not fully rational) behavior?

Starting from the early days of the ethical debate on nudging, one popular answer to this question among AO supporters has been as follows: In order for an agent to count as a fully rational (or so the critic maintains), her thoughts and actions have to be based on – in the sense of being adequately responsive to – *reasons*. However, the critic points out, most nudges, by their very nature, manifestly fail to count as genuine reasons for either believing or doing anything. It

[15] Cf., e.g., Bovens (2009), Hausman & Welch (2010), Grüne-Yanoff (2012), Saghai (2013), MacKay & Robinson (2016), and Viale (2022).

[16] Cf., e.g., Engelen (2019), and Thaler & Sunstein (2021).

[17] Cf., e.g., Barton & Grüne-Yanoff (2015).

[18] Cf., e.g., Schmidt (2019). See Todd & Gigerenzer (2012) for an extensive discussion and defense of the notion of *ecological rationality*.

[19] Cf., e.g., Hansen & Jespersen (2013), and Schmidt (2019).

follows, she concludes, that nudged choices are intrinsically irrational ones. This line of reasoning appears to be central to AO supporters' anti-nudge case. Explicit endorsements thereof are already to be found in early criticisms of nudging, such as the one leveled by Bovens (2009). Nudged choices, in his view, would fall short of rationality as the psychological mechanisms that choice architects rely on do not count as reasons. In particular, insofar as the mechanisms that guide our actions would allegedly fail to constitute genuine reasons for acting, upon being nudged we would no longer be fully in control of our behavior.[20] A similar worry has more recently been expressed by Viale (2022), according to whom nudgers would be culpable of steering us toward behaviors that are not grounded on reasons.[21] If this line of reasoning were on the right track, then – quite regardless of our motives for doing so – allowing nudgers to guide us in a given direction by intervening on our decisional environment would indeed be tantamount to willfully engaging in a form of epistemically irrational behavior. As we shall see in the next section, however, the idea that allowing to be affected by nudges would be something intrinsically irrational can be shown to rest on a woefully enduring, and yet deeply misleading myth about intellectual autonomy, and it should therefore be resisted.

2.3. Nudges as reasons

According to AO, as we just saw, nudges would fail to constitute genuine reasons for thinking or acting. If that were the case, it would indeed be irrational to let them steer our choices. As AO supporters – just as we do – regard rational agency as a central human value, they maintain that reliance on nudges in policy should be opposed on moral grounds. In this section, I will argue that AO, while motivated by sharable concerns, rests on a deeply implausible understanding of intellectual autonomy. Once this confusion is dispelled, I maintain, nudges can be seen as not just fully compatible with our rational agency, but also as an extremely valuable way to boost it. Allowing nudgers to affect our decisions, in my view, can often be not only fully rational, but also, epistemically speaking, the best thing to do. Let us then begin by engaging more closely with AO's take on intellectual autonomy.

The kind of intellectual autonomy that AO supporters seem to typically have in mind in opposing the use of nudges is usually referred to in the epistemological literature as *intellectual individualism* (henceforth, II).[22] II incarnates an austere ideal of complete epistemic self-reliance. It is the view that in order to count as fully rational, an epistemic agent should depend on no one but herself in conducting inquiry – i.e. constantly strive to reach correct answers to the myriad of questions that confront us every day by counting solely on the powers of her own, unaided cognition. On this stern view of our epistemic duties, in particular, as it has been pointed out, a fully rational agent could only be someone who refuses to take anyone's word for anything, and relies exclusively on her own cognitive and inferential powers.[23] Now, if II were indeed what we should take intellectual autonomy to consist in, very few people, I suspect, would keep regarding it as an epistemic virtue worth aiming at. As many epistemologists have long acknowledged, however, II

[20] Bovens (2009: 209).
[21] Viale (2022: 14).
[22] Cf., e.g., Matheson (2024).
[23] Fricker (2006: 225).

is a myth.[24] As a normative ideal, it is simply not in line with the actual way in which a deeply social phenomenon such as human higher cognition has been amply shown to normally develop and work.[25]

The most straightforward way to demonstrate that II is just off perhaps consists in drawing on a common epistemological distinction – i.e. the one between *direct* and *indirect* reasons that an agent may possess for believing as she does. The former kind of reasons are normally so called because, contrary to the latter, they *directly* support a given claim. Consider the following situation. As I sit across from Sarah in our office, my perceptual experience of seeing her working at her desk is a *direct* reason that I have to believe that she is in her office. Suppose now that John calls me on the phone inquiring about Sarah's whereabouts. When I tell him that Sarah is in her office, John acquires an *indirect* reason to believe this claim by taking my word for it. Indirect reasons, then, are reasons about reasons. It has therefore become common to refer to them as *higher-order reasons* or *higher-order evidence*.[26] By telling John that I am looking at Sarah as we speak, I provide him with a higher-order reason R_2 – i.e. my testimony – to believe that I have a first-order reason R_1 – i.e. my perceptual experience – to believe as I do. Provided that John has access to independent reasons (R_3 … R_n) to trust me on the matter, he thereby acquires an indirect reason (R_2) to believe that Sarah is in her office, and, according to most epistemologists, R_2 counts as a fully legitimate epistemic reason for him to believe that much.

It is a plain fact that most of the beliefs that allow us to navigate our physical and social world are based on indirect reasons. Importantly, many such beliefs are far more consequential than the one used in my toy example. Take, for instance, the belief that anthropogenic climate change is occurring. It would be hard to deny that the vast majority of people who hold it (myself included) do so based on indirect reasons. Whereas climate scientists are indeed rightly expected to possess or at least have access to direct reasons to believe this claim, the normal division of epistemic labor makes it the case that most of us will be far better off *deferring* to their expertise when it comes to figuring out the truth about matters, such as the present one, that are far beyond the reach of our own knowledge and intellectual abilities. The most effective way to settle the issue, from where we (epistemically) stand, would clearly be to trust the experts' testimony and 'take their word for it'. In similar situations, that is, scientists will in effect play the role of what Dogramaci (2012) refers to as *epistemic surrogates* – i.e., individuals or groups to whom we can outsource inquiry on certain matters, and by trusting whose testimony we can acquire (indirect) reasons for holding various beliefs.[27] Note, however, that if II were indeed a plausible understanding of intellectual autonomy, then it would follow from my last example that, for the vast majority of people, it is less than fully rational to believe that anthropogenic climate change is indeed occurring – a hardly palatable conclusion. As I would like to suggest, however, absent compelling reasons to distrust the testimony of the relevant experts, there is nothing epistemically inappropriate or agency-threatening with this kind of intellectual outsourcing, and it is perfectly rational for said majority to believe as they do. It would appear, then, that intellectual autonomy is much more plausibly construed as an eminently relational epistemic virtue. Once so construed,

[24] Cf. Matheson (2024: ch. 4, fn. 6).

[25] Levy (2022) makes a convincing case for the claim that what has long prevented the western epistemological tradition from fully acknowledging this point is its almost exclusive focus on individual cognition.

[26] Cf. Horowitz (2022).

[27] Cf. Dogramaci (2012: 524).

I shall now argue, our rational agency can be seen as wholly compatible with the use of nudges as policy instruments.

The idea that it can be fully rational to let nudges affect our daily choices has perhaps been most forcefully defended in the work of Neil Levy.[28] His general take on nudging is premised on the empirically well supported assumption that unaided individual cognition is highly unreliable and that, as a consequence, most of our intellectual achievements are manifestly due to our social interactions with other past and present epistemic agents. Insofar as knowledge acquisition in our species heavily relies on social learning[29], I wholeheartedly agree with Levy that, absent ample opportunities to defer to the expertise of others on a plurality of matters of either theoretical or practical import, we would be "epistemically at sea."[30] Most of our beliefs, as I put it above, are indeed based on indirect reasons – i.e. reasons we acquire by relying on the testimony of others. In my previous examples, for instance, John acquires an indirect reason to believe that Sarah is in her office by trusting my testimony to that effect, and most people today have access to indirect reasons to believe that anthropogenic climate change is occurring by trusting scientists' testimony to that effect. As Levy points out, however, our social environment is one in which testimony is not always explicit, but frequently of an *implicit* sort. Indeed, the kind of evidence that humans automatically and systematically rely on in navigating their social world is often not delivered by means of assertions, but rather conveyed by various cues disseminated throughout our environment. To exemplify, the fact that news A appears on the front page of a newspaper, whereas news B gets relegated to its last page, can easily be seen and treated as a form of implicit testimony, as it encodes information about the further fact that, evidently, the majority of readers deem news A to be far more important that news B. This further fact, in turn, may constitute higher-order evidence in favor of the proposition that news A *is* indeed more important than news B, and, to the extent that people respond to it accordingly, it clearly counts as a reason for thinking and acting.

As Levy convincingly argues, nudges are just another form of implicit testimony – i.e. social cues that agents respond to by treating them as higher-order evidence in favor of a given course of action.[31] Their impact on our choices, in particular, would be due to their providing implicit recommendations to the nudgee. Contrary to what AO supporters maintain, then, nudges *are* in effect ways of providing us with *reasons* to engage in specific behaviors, and, to this extent, they are wholly compatible with our rational agency. Let us hence refer to this stance as the *reason-giving view* of nudges (henceforth, RGV). An indication that RGV is on the right track, in my view, consists in the existence of empirical evidence suggesting that, as a matter of fact, this is precisely the way in which people tend to perceive a type of nudge that will come to play a central role in what follows – i.e. defaults.[32] As we shall see, people do often interpret defaults as social cues conveying useful information, and resist their persuasive pull accordingly – i.e. absent discounting cues such as, e.g., lack of credibility or self-serving motives on the part of the choice architect.

---

[28] Cf., in particular, Levy (2022).

[29] Cf., e.g., Boyd et *al.* (2011).

[30] Levy (2022: xvii).

[31] Cf. Levy (2022: 139). As Levy points out, not much hangs on our willingness to call 'implicit testimony' the kind of information transfer enabled by such cues. What counts is rather that the information conveyed is processed by its recipients as evidence (Levy 2022: ch. 2, fn. 6). As it will become clear in what follows, I agree with him on this score.

[32] Cf., e.g., McKenzie et *al.* (2006), and Sunstein (2016a: 169-70).

Similar findings, I take it, suggest that the currently common picture of defaults as underhanded compliance techniques that bypass our rational agency by exploiting our cognitive shortcomings may in fact be widely off the mark. Let me hence move on and put both AO and RGV to the test by taking a closer look at a very promising type of green nudges – i.e. energy defaults.

2.4. Automatically green

It is high time for me to acknowledge what I regard as a significant limitation of the above considerations. The problem, as I see it, is that both AO and RGV target the use of nudges in general. This raises a crucial methodological issue: does it really make sense to morally object to or defend the use of nudges *in general*? At various points in his book on the ethics of nudging, Sunstein flags the risk of falling into what he fittingly dubs the 'trap of abstraction' – i.e. to casually assume that the ethical assessment of a specific nudge will straightforwardly apply to all types of nudges.[33] I see at least two reasons for holding that this would indeed be a bad idea. Firstly, not all nudges pursue the same ends – a fact that clearly bears on their moral status. Recall, for instance, that according to the distinction between *pro-social* and *pro-self* nudges, green nudges – whose moral credentials I am here considering – are not primarily aimed at improving the welfare of the nudgee. It follows, as it has been repeatedly observed, that they cannot be fully justified simply on paternalistic grounds.[34] The second, and possibly more weighty reason, has not so much to do with nudges' ends, but with their cognitive underpinnings. In this regard, to the extent that different types of nudges leverage different psychological mechanisms, it seems only fair to doubt that any given objection to or defense of their use will apply with equal force (or even, indeed, at all) to each and every type of nudge. All in all, it seems far more plausible to expect that nudges relying on different mechanisms will display distinct ethical profiles, and will hence call for independent considerations and assessments.[35]

How do we then steer clear of the trap of abstraction? Sunstein's own advice is to bring the above considerations about nudging and rational agency into close contact with concrete practices[36], and this is precisely what I intend to do. In what follows, I will hence focus on a specific type of green nudges targeting energy-related CRBs – i.e. energy defaults. This choice is motivated by the following two considerations. On the one hand, defaults are currently among the most effective nudges in choice architects' tool box.[37] In particular, as we shall presently see, they have proved very promising in encouraging greener energy provision.[38] On the other hand, defaults are also widely regarded by nudge critics as one of, if not *the* most ethically problematic type of nudges, as they at least prima facie appear to exclusively target our System 1, thereby entirely

[33] Cf., e.g., Sunstein (2016a, 15-16, 26).
[34] Cf., e.g., Nagatsu (2015), Oliver (2013, 2017), Schubert (2017), and Congiu & Moscati (2022).
[35] Interestingly, some surveys suggest that people do not care, nor have any specific views, about nudges *in general*. Their attitudes seem to rather depend on whether they approve of the goals and workings of particular types of nudges. Cf. Sunstein (2016a: 118).
[36] Cf. Sunstein (2016a: 16).
[37] Cf., e.g., Sunstein & Reisch (2014).
[38] Cf., e.g., Sunstein & Reisch (2013, 2014), Sunstein (2016a: Ch.7), and Schubert (2017).

bypassing our conscious reasoning and deliberation.[39] As a consequence, if AO can be shown to be ill or insufficiently motivated in the case of defaults, then this should come as good news for nudgers, as it should arguably enhance the prospects of morally justifying other, admittedly less controversial, types of nudges. Let us then take a closer look at energy defaults.

The following point, I take it, stands hardly in need to be argued for. It is today in both present and future humanity's best interest to cut down on the burning of fossil fuels, such as oil, coal, and gas, by increasing the reliance on renewable energy sources such as water, wind, biomass or the sun. It is no wonder, then, if the International Energy Agency has long been calling for a coordinated governmental action on this front, and the European Commission has long followed suit by promoting the use of renewable energy sources.[40] In short, we currently face the pressing need to make our energy use sustainable, and an important step toward achieving this goal is to encourage households to decrease their manifest overreliance on grey energy – i.e. to increase enrollment rates in greener energy provision. The good news, in this regard, is that, due to the way in which electricity markets, in many countries, are currently organized, consumers today can not only know how the electricity they use is produced, but also act on this knowledge by electing greener energy tariffs.[41] It would appear, however, that this is much easier said than done. Surveys indeed show that, in spite of their explicitly reported environmental values, only very few people actually end up (literally!) putting their money where their mouth is. When polled, people typically manifest a strong preference for greener energy. Yet, unfortunately, they hardly ever act on it.[42] This lack of correlation between environmental attitudes and behaviors, as we saw in the first chapter, has long been experimentally investigated and theoretically addressed by social psychologists, who usually refer to it as the *Value-Action Gap*.[43] As we shall see, however, energy defaults happen to constitute a very powerful tool at our disposal to bridge the gap in the case of greener energy consumption.

Defaults are options in our choice architecture that prevail when we do not actively choose. The so-called *default effect* consists in the observed, strong tendency of people not to change an option selected by someone else, even in cases in which the costs involved in so doing would be very low. As applied to the case of energy consumption, this means that households will tend to stick with whatever electricity tariff they are presented with as a default option by their suppliers. In light of defaults' large impact on other human behaviors[44], it is therefore reasonable to expect that the enrollment rates for green, as opposed to grey, energy provision will themselves increase or decrease as a function of whether customers are presented with either green or grey default options respectively. One of the most striking illustrations of this prediction comes from two

[39] Some authors distinguish Type 1 from Type 2 nudges, according to whether they target our System 1 or our System 2 respectively. Whereas Type 2 nudges would trigger conscious reasoning and deliberation, Type 1 nudges would rather bypass it entirely. Cf., e.g., Hansen & Jaspersen (2013), and Sunstein (2016b). As I shall argue, however, evidence suggests that energy defaults are not correctly described as nudges that only engage our System 1.

[40] Cf. Pichert & Katsikopoulos (2008: 63).

[41] Cf., e.g., Bird et *al.* (2002), and Clark et *al.* (2003).

[42] Cf., e.g., Bamberg (2003), Clark et *al.* (2003), and Rowlands et *al.* (2004).

[43] Cf., e.g., Kollmuss & Agyeman (2002), Gifford (2011), Lacroix & Gifford (2017), and Pirni (2023).

[44] Johnson & Goldstein (2003), for instance, have shown that consent rates for organ donation vary very widely across European countries according to whether the country in question adopts either an opt-in or an opt-our policy on the matter. At the time of their study, an opt-out country such as Austria had reached an astonishing 99% of potential organ donors, whereas in Germany, an opt-in country, organ donation rates were down at 12%.

justly famous field studies conducted in Germany, and showing that, in real-world, as opposed to laboratory-settings, energy defaults are indeed remarkably effective.[45]

The first case is the one of Schönau, a small town in the Black Forest, where in 1997 a citizens' initiative which had long been campaigning in favor of taking over the local electricity grid to establish a green energy supply finally managed to buy off its previous owner, and to promote the use of solar energy by making it the default option on the local market. In spite of a significant initial reluctance on the part of the citizens[46], by 2006 nearly every customer had remained with the green default.[47] The second case is that of Energiedienst GmbH, a German company that in 1999 mailed letters to its customers advertising three new energy tariffs. Customers who did not reply would be enrolled in a default green tariff – this time the renewable energy source was waterpower. Whereas the new default was slightly cheaper than the old tariff, the other two new tariffs were a grey (about 8% cheaper), and an even greener (about 23% more expansive) one. The effect of this marketing strategy was that after two months, about 94% of the customers, although having been given the chance to switch to the cheaper tariff, had remained with the green default option. What makes these two cases particularly impressive is that, at around the time they took place, the share of people participating in green electricity programs in other parts of Germany where grey energy defaults were in place was down to 1%.

Even factoring in the well-known methodological limits of studies targeting naturally occurring phenomena – such as, e.g., the obvious lack of experimental control[48] – similar success stories, as it has been noted, testify to the extraordinary power of defaults.[49] At least from a behavioral perspective, then, there is little doubt that this power can be profitably harnessed to foster the adoption of CRBs aimed at decreasing the use of grey energy sources.[50] Especially at times, such as the present one, where the stakes have grown very high for both living and future humanity, there hence seems to be only one sensible advice that behavioral scientists should give to policy makers – i.e. implement energy defaults as widely as possible! As we saw above, however, not everybody feels the same way about the prospect of systematically relying on green nudges in policy. Let me then go back to normative matters, and draw on the considerations developed in section 2.3 to make my final case for energy defaults.

---

[45] Pichert & Katsikopoulos (2008).

[46] Indeed, what makes this case particularly telling is that the proposal to take over the local electricity grid, once put to the vote, was only accepted by a very close margin – i.e. 52% vs. 48% (the turnout was approximately 90%).

[47] Interestingly, further details of this case suggest that demographic variables, such as e.g. socio-economic status, cannot plausibly be appealed to in order to explain this effect. Schönau, for instance, was at the time politically dominated by conservatives, notoriously not in favor of pro-environmental policies.

[48] I hasten to add the Pichert & Katsikopoulos (2008) compensate for this by presenting two additional laboratory experiments that offer further support to their general hypothesis, according to which in cases where green electricity is the default, most people will use it. Cf. Pichert & Katsikopoulos (2008: 67-70).

[49] Cf. Sunstein (2016a: 165).

[50] In the United Kingdom, for instance, the Department of Energy and Climate Change already acknowledged this point in 2011 by highlighting the effects of defaults in a report on policies aimed at reducing energy use. Cf. Smith et *al.* (2013: 168).

2.5. Energy defaults as reasons for climate action

For reasons already discussed, it seems unlikely that any moral objection to the use of nudges in policy will apply with equal force to each and every type of nudge. By parity of reasoning, however, any given defense of nudges' moral credentials will likely result more or less convincing, depending on which type of nudge it is applied to. As a consequence, to avoid the 'trap of abstraction', I decided to restrict my attention to a specific type of nudge. It is hence time to see how both AO and RGV fare with respect to energy defaults. According to AO, as you will recall, nudges in general would seriously undermine our rational agency – i.e. the ability to make autonomous decisions and act on them – by tapping on irrational decisional processes. More to the point, they would fail to constitute genuine reasons for action, and nudged choices should therefore be considered intrinsically irrational. As we have seen, however, AO seems driven by an implausible understanding of intellectual autonomy. Once this epistemic virtue is properly construed, I claimed, nudges appear not just fully compatible with our rational agency, but also as an extremely valuable way to boost it. I also sided with supporters of the mischaracterization response to AO in claiming that the objection relies on an incomplete understanding of nudges' psychology. Let us then see how all this applies to the case of energy defaults.

The ethical implications of defaults, as it has been wisely observed, cannot be objectively assessed without a theory of why default effects exist at all.[51] In this regard, there are currently four dominant theoretical explanations of defaults' effectiveness on the market, each pointing to a different cognitive mechanism. The four main psychological factors that have been put forward as causally responsible for default effects are *inertia*, *loss aversion*, *guilt*, and *implicit recommendation*.[52] As you may suspect by now, the last factor is the one that I am here interested in considering more closely. Indeed, as anticipated above, research supports the idea that people tend to interpret defaults in general as social cues conveying an endorsed or recommended expert opinion, and they respond to them accordingly. McKenzie and colleagues (2006), for instance, provide evidence that, in the domains of organ donation and retirement plans, people perceive default options as implicitly suggested courses of action.[53] Their findings indeed suggest that, by selecting a given default, policy makers would be sending out – or, as they put it, 'leak' – information about their beliefs or attitudes toward the available options; information that decision makers would in turn 'absorb'.[54] Importantly, their results clearly show that, contrary to the standard narrative, people are usually *aware* of how a given default affects themselves and others.[55] This last point is crucial to my case, and it therefore requires further elaboration. From the present perspective, indeed, there is one important aspect that sets the implicit recommendation view apart from other existing accounts of default effects – i.e. its explicit acknowledgement of the *metacognitive* component involved in many decisional processes. Let us then take a closer look at this feature.

In unpacking AO, we agreed to understand a nudged choice as the behavioral output of an often automatically-triggered decisional process occurring within the context of a choice architecture

[51] Smith et *al.* (2013: 169).
[52] Cf., e.g., Sunstein (2016a:169-173).
[53] McKenzie et *al.* (2006).
[54] Cf. McKenzie (2004).
[55] Cf. McKenzie et *al.* (2006: 417).

intentionally designed to promote a given behavior. This standard characterization, I believe, is broadly correct, as far as it goes. I also believe, however, that in addressing more closely the psychology of nudging one should not lose sight of the obvious fact that, quite often, in making decisions and judgements, we are not just idle bystanders to our cognitive lives. This peculiar aspect of human cognition, I suggest, while having important implications for my case, has not yet received enough attention within the nudging literature. As I observed above, within this literature defaults are typically pointed at as paradigmatic examples of Type 1 nudges – i.e. nudges that affect our behavior by entirely bypassing conscious reasoning and deliberation. It would appear, however, that once we allow metacognition to enter the picture, things start looking remarkably different, and the currently predominant view of defaults begins to falter. Perhaps unsurprisingly, as we shall see, this last point has long made its way into a field of social science traditionally characterized by a less abstract, and more hands-on approach to its subject matter – i.e. consumer research.

In 2002, Peter Wright envisioned a promising new area of consumer research that he fittingly named 'behavioral marketplace theory'.[56] In outlining its main object of study, he first introduced the notion of *marketplace metacognition* (henceforth, MM), which he conceived of as a pragmatic form of social intelligence whose proper domain of application were marketplace thinking and behavior, and that individuals would keep developing over their life span. MM, in his view, should therefore be understood as encompassing an increasingly growing set of mental 'routines' and 'contents' that we normally bring to bear on our everyday marketplace transactions in order to make good decisions.[57] While very general in scope, Wright's ideas bear on the present discussion as they happened to prove remarkably useful in advancing research of defaults' psychology.

In the wake of Wright's seminal essay, Brown and Krishna (2004) decided to experimentally challenge the received wisdom about defaults by looking at them through the lens of MM.[58] According to the explanations they were up against, defaults would affect our choices by exploiting our processing limitations.[59] Let us then collectively refer to their polemical target as the 'deficit accounts'. In order to distance themselves from such accounts, Brown and Krishna put forward a *metacognitive account* of defaults, and presented the results of two studies meant to support it. Building on previous work in consumer research[60], they started from the assumption that people treat defaults as 'carriers of meaning' that provide relevant information about both marketers behavior and product value. What they found is that, in general, MM plays a crucial moderating role with respect to the way in which people respond to defaults. In particular, their results show that the size and direction of a default effect will largely depend on how a given consumer's MM affects her interpretation of the default at hand. If the consumer interprets a given default option as biased in favor of the marketer, for instance, she will be more likely to respond to the perceived manipulation by discounting it.

---

[56] Cf. Wright (2002: 678, 681). Importantly, for my purposes, at several points of his seminal essay, Wright makes clear that a defining goal of the new research area should consist in letting its empirical finding inform programs aimed at fostering an egalitarian and research-based marketplace education. Cf. Wright (2002: 682).

[57] Cf. Wright (2002: 677).

[58] Brown & Krishna (2004).

[59] Brown and Krishna take their polemical targets to fall into two overall categories, which they respectively refer to as 'attention-based default effects' and 'default effects due to processing distortions'. Cf. Brown & Krishna (2004: 529). These two categories essentially overlap with two of the accounts mentioned at the beginning of this section – i.e. inertia, and loss aversion. Guilt-based accounts are not considered in their study.

[60] Cf. Prelec et *al.* (1997), Wernerfelt (1995), and Briley et *al.* (2000).

Leaving details aside, three points emerge from Brown and Krishna's work that are here worth emphasizing. First of all, their metacognitive account dovetails nicely with other existing evidence suggesting that defaults' effectiveness is indeed significantly affected by the nudgee's level of trust in choice architects, and by her expertise or experience with the options in question.[61] The second point is that, as the authors discuss at various points, many effects predicted by their account and observed in their study simply cannot be accommodated within the frameworks provided by deficit accounts.[62] Their metacognitive account, then, would appear to outcompete its rivals in predictive power, and, to the extent this theoretical virtue is commonly regarded as a pivotal one in theory choice, this should count as a reason to regard it as closer to the truth about defaults' psychology. The third, and crucial point is that Brown and Krishna's findings appear to call into question defaults' alleged status as paradigmatic Type 1 nudges. Their results indeed strongly suggest that, far from reaping the benefits of our cognitive limitations, defaults can make consumers more skeptical and alert by engaging their MM.[63] In other words, contrary to the received wisdom, defaults would not at all bypass our conscious reasoning and deliberation – they would rather boost it by triggering inferential processes whose conclusions can be either in favor or against the implicitly recommended option.

While of course not conclusive, I take the above considerations to count as robust evidence in favor of RGV, as this view of nudges applies to the case of energy defaults. As I hope to have shown, once metacognition is taken into due account, energy defaults *are* indeed best seen as a form of implicit testimony – i.e. subtle features of our decisional environment that provide us with indirect, higher-order *reasons* in favor of a given option. It follows that, contrary to what AO supporters have been alleging, responding to these social cues cannot be plausibly equated with engaging in a form of irrational behavior, but should rather be seen and treated as an utterly normal, and epistemically praiseworthy kind of intellectual outsourcing. Indeed – absent valid reasons for distrust or a (typically uncommon) solid expertise in energy matters – it appears fully rational to think of choice architects as profitable epistemic surrogates, and to let them steer our choices by selecting a given default option in our stead. Whenever they are exposed to default options, according to Smith and colleagues (2013), consumers would unavoidably end up handing some of their independence over to the market, and their intellectual autonomy would thereby be diminished.[64] As I understand their reasoning, the authors are here drawing a general inference that indiscriminately ties every instance of intellectual outsourcing to a corresponding decrease in intellectual autonomy. Note, however, that this inference crucially relies on a tacit premise. It can only appear self-evident to someone who is (more or less consciously) already committed to the truth of a previously discussed myth – i.e. intellectual individualism. Once this hopeless way of thinking about autonomy is off the table, the inference loses much of its initially intuitive appeal, and it actually appears much less convincing. To conclude, I take the above considerations to have shown that AO-fueled worries ultimately rest on an implausible understanding of intellectual autonomy, coupled with an incomplete grasp of defaults' cognitive underpinnings.

[61] Cf., e.g., McKenzie et *al.* (2006), and Löfgren et *al.* (2012). McKenzie and colleagues take their results to provide evidence for both claims. Cf. McKenzie et *al.* (2006: 419). Löfgren et *al.* (2012) provide evidence that experience with CO_2 offsetting programs attenuates the effect of corresponding defaults on environmental and resource economists.

[62] Cf., e.g., Brown & Krishna (2004: 530, 532, 536, 537).

[63] Cf. Brown & Krishna (2004: 531).

[64] Cf. Smith et *al.* (2013, 163).

Once properly construed, I argued, energy defaults cease to represent a threat to our rational agency – they do not undermine our intellectual autonomy, but rather constitute a valuable opportunity to more fully exercise it. As a consequence, their use should be regarded as morally legitimate, and policy makers should fully support their wide implementation as a promising and profitable means to foster CRBs aimed at reducing energy consumption. Energy defaults provide us with genuine reasons for climate action.

3. The Case for Nudged Beliefs

3.1 Practicing what we preach

Natural born influencers, as I noted in section 1.1 above, are bound to also be natural born psychologists.[1] An ability that every cub of our species needs to develop in order to effectively navigate the social world is indeed to make sense of other individuals' actions. Consider the following question: "why did Clara stealthily sneak behind the bush?" Very early in life we begin to rely on relevant background knowledge in order to answer similar questions by attributing various mental states, among which *beliefs* feature prominently, to people – e.g. "Clara sneaked behind the bush because she is playing hide-and-seek with her brothers, and she *believes* that Giuliano is hiding there." This general tendency, of course, does not end with childhood, but accompanies us throughout our adult lives. Think again about the persuasion data discussed in section 1.2. Why was Hovland's group so interested in studying people's attitudes toward the war? What made it so urgent to the Yale team, and especially to the people in Washington who had hired them, to reach a better understanding of the factors that made the formation of those opinions more likely? Or think about today's advertising and political campaigning – currently, in the U.S., a multi-billion-dollar industry. Why would, e.g., Mr. Trump or Mr. Zuckerberg care in the least about what most Americans like or think?[2] It seems a fair bet to assume that, while the immediate target of their expansive advertising and campaigning activity are indeed people's opinions, the ultimate goal of those activities is rather to alter, respectively, people's voting or consuming *behavior*. Both advertising and campaigning, that is, are clearly premised on the widely shared assumption that altering the way in which people think and feel about various issues will eventually bring about a change in the way they act. Now, the above observations, I take it, strongly suggest that our brains are wired in a way such as to find the following thought intuitively compelling:

> Beliefs *shape* behaviors, and can hence be relied on to *explain* and *predict* them.

As intuitions about such matters are likely to be a dime a dozen, however, it seems fair to ask: is this proposition actually true? The issue is a complex one, and psychologists have been addressing it systematically for almost a century. On the one hand, starting from the 1950s, the so-called *principle of cognitive consistency* has played a central role in most psychological theorizing.[3] In line with this guiding principle, as I noted in section 1.4, early research on attitudes simply assumed

[1] The idea that homo sapiens would be best characterized as *homo psychologicus* has been defended, for instance, in Humphrey (1984), according to whom human beings would be 'born psychologists' as they very early begin to manifest an inborn tendency to account for people's behaviors in terms of their mental states.

[2] Granted, they may be interested in finding it out in order to design better policies or to better meet the needs of their customers respectively, but I suspect that this is hardly the main drive behind their inquiring activity.

[3] Two sourcebooks on cognitive consistency are Abelson et *al.* (1968), and Gawronski & Strack (2012). As we shall see in the next section, the notion plays a central role in what is arguably social psychology's most influential theory – i.e. Festinger's (1957) theory of cognitive dissonance. Cf. Kruglanski et *al.* (2018) for a recent attempt to question the existence of a universal human need for cognitive consistency.

that people's actions and behaviors will normally be consistent with their opinions and attitudes. So much so, indeed, that early theorists often included behaviors in the very definition of attitudes – variously characterized as, e.g., "implicit behaviors', 'implicit responses', or 'acquired behavioral dispositions'.[4] On the other hand, this state of affairs had not prevented researchers from raising occasional doubts concerning its validity. A milestone, in this regard, is represented by a famous field study on racial prejudice in the United States conducted at the beginning of the 1930s by Stanford sociologist Richard LaPiere.[5] The origin of this study is worth reporting.

At the time, anti-Chinese prejudice was high in U.S. On one occasion, LaPierre happened to be traveling by car with a young Chinese couple (a student and his wife), and noticed that, contrary to his initial worries, the manager of the hotel where they stopped for the night accommodated them "without a show of hesitation". Surprised by the unexpected event, the sociologist decided to phone the hotel after two months in order to inquire whether they would by any chance be willing to host "an important Chines gentleman". The answer was *no*. That single answer, in LaPierre's own words, "aroused my curiosity and led to this study."[6] As a consequence, from 1930 to 1932 LaPierre traveled by car twice across the United States with his young (and, by today's standards, very cooperative!) student and his wife. Out of the 251 hotels and restaurants where they stopped, only *one* refused to serve them. After six months – a time supposed to allow his subjects to forget about their past hospitality – he mailed questionnaires to the owners of the venues they had visited. One of the questions, in particular, read as follows: "Will you accept members of the Chinese race as guests in your establishment?" As expected, 92% of the 128 replies that he received back, gave the same answer: *no*. Here is the conclusion that LaPierre himself drew from his study:

> "On the basis of the above data it would appear foolhardy for a Chinese to attempt to travel in the United States. And yet, as I have shown, actual experience indicates that the American people, as represented by the personnel of hotels, restaurants, etc., are not at all averse to fraternizing with Chinese within the limitations which apply to social relationships between Americans themselves."[7]

Contrary to our lay intuitions then – and, crucially, contrary to an unqualified version of principle which at the time informed most psychological research – the message that seemed to come from LaPiere's study, although somewhat disheartening, was nonetheless pretty clear – i.e., the vast majority of people do not at all seem to practice what they preach.

How did psychologists initially react to the "bad" news? Well, certainly not by ignoring it, yet also, by and large, neither by jettisoning the idea of an overall attitude-behavior consistency. Things were however bound to change within a couple of decades. Indeed, starting from the 1960s, the existing evidence for an overall correlation between our attitudes and behaviors began to be systematically challenged by a series of meta-analytic reviews.[8] While detailing such developments would fall beyond the scope of the present discussion, what matters for our purposes is rather the upshot of subsequent research on the matter. In this regard, as it is often

[4] Cf. Hovland et *al.* (1957), Doob (1947), and Campbell (1963), respectively.
[5] Cf. LaPiere (1934).
[6] *Ibid.*, p. 232.
[7] *Ibid.*, p. 234.
[8] Cf., e.g., Wicker (1969).

the case in science, what became increasingly clear is that the above principle of cognitive consistency is far too general as it stands, and it hence needs to be more carefully qualified. Most importantly, evidence pointed to the fact that the complex relation between attitudes and behaviors is usually moderated by a vast array of other variables, such as, e.g., a given attitude's strength, the particular way in which it was originally formed, or else the specific reasons why the individual may hold it. As a consequence, we should normally expect people's opinions to be more reliable as a guide to their behaviors under certain conditions, and less so under others.[9] The collective recognition of this fact brought about a shift in theoretical focus, and a corresponding reconsideration of the goals of experimental research. As of today, attitudes are generally regarded to be fairly reliable guides to our behaviors in many domains, and the central question addressed by social psychologist is no longer *whether* our opinions predict our actions, but rather *when* they do – i.e. under which conditions are we more likely to act in accordance with our beliefs, or, as I put it above, to practice what we preach? Current research, in other words, aims at locating factors that can be shown to affect – i.e. increase or decrease – attitude-behavior consistency. Some of them will be considered in due course. Before doing so, however, it will be necessary to interleave our discussion with a brief terminological digression.

The present dissertation is premised on the methodological assumption that – as much excellent, past and present philosophical work stands to show – philosophical and psychological results can often helpfully illuminate each other. As anyone who ever tried to meet this challenge by stepping outside of her disciplinary comfort zone can testify, however, this is unfortunately much easier said than (properly) done. One problem, in our case, is that both philosophers and psychologists make abundant use of the word 'attitude'. The rub, alas, is that they happen to use this term in importantly different ways. Here is how Peter Carruthers – a philosopher of mind who not only shares, but also masterfully implements the above methodological assumption – has summarized this state of affairs:

> "In *psychology* an attitude is, roughly, a disposition to engage in evaluative behavior of some sort. Thus, one has an attitude towards a political party, or the morality of abortion, or the permissibility of the death penalty. But one doesn't (normally) have an attitude towards the date of one's own … birth, or to the fact that whales are mammals.
> In *philosophy* … in contrast, an attitude can be any kind of standing thought … that has a conceptual or propositional content. … Hence knowing, or recalling, that I was born in June are propositional attitudes. Believing, or judging, that whales are mammals are propositional attitudes. And so, too, are wanting, hoping, fearing, supposing, or being angry that the next president will be a Republican."[10]

As it is to be expected, this difference in use is liable to create several misunderstandings when one intends to bring empirical results from psychology to bear on philosophical issues and vice versa – as the present dissertation aims to do. Let me then try to steer clear form this risk by quickly introducing a few basic notions, and the related terminology, that are presupposed in social psychologist's talk of 'attitudes'.

---

[9] Cf., in particular, Kraus (1995), whose meta-analysis revealed a significant impact of attitudes on behavior across 88 studies.

[10] Carruthers (2011: Preface, xiii).

In maximally general terms, the psychological construct that psychologists refer to as an *attitude* is simply a categorization of a given *stimulus object* (often called the *attitude object*) along an evaluative dimension – and it will accordingly vary in its degree of positivity or negativity – i.e. its *valence*. An attitude object can in turn be characterized as pretty much anything about which we can form an evaluation, such as, e.g., material objects, people, places, experiences or issues. Attitudes are then standardly taken to significantly differ from one another in terms of their *structure*, *strength*, and *function* – three aspects that, as we shall see, play a crucial role when considering their relation to behaviors. With respect to their structure, attitudes have traditionally been conceived as having three main components[11] – i.e. a *cognitive*, an *affective*, and a *conative* one, respectively encompassing our beliefs, our feelings, and our behavioral intentions and actions toward the object. The *strength* of a given attitude is instead usually defined by two markers, namely *durability* – i.e. its stability over time and resistance to change under persuasion or new information – and *pervasiveness* – i.e. its impact on our behavior and on the formation of other attitudes. In this regard, researchers have so far identified several factors that are now commonly taken to have an impact on attitude strength, some of which will come up below, as they have been found to affect attitude-behavior consistency as well. How about the last of the three aspects mentioned above – i.e. the attitude *function*? A convenient entry point here consists in taking a step back by asking why we form and hold attitudes at all, or, in other words, what purpose they serve. Now, according to the currently prevailing, functional approach to this issue, the answer is that such mental states help us navigate our physical and social world by responding to various needs. Their *general* function, that is, would be to fulfill basic human motivations. Their *specific* function, on the other hand, will vary according to which one of such motivations the holding of a given attitude is intended to fulfill for a given person and in a given situation. Research on this front has singled out the following five main motivations (and corresponding attitude functions), which can mutually interact with each other[12]: (1) forming accurate beliefs about the world (*epistemic*), (2) avoiding dangerous situations (*utilitarian*), (3) expressing our values (*value-expressive*), (4) fitting in with a given reference group (*social-adjustive*), and (5) protecting our self-esteem (*ego-defensive*). Again, some of these functions will come up below, as they have also been found to affect attitude-behavior consistency. In light of the above, the class of mental states that philosophers refer to as *beliefs* – or, more in general, *doxastic attitudes* – can, and will be, seen and treated in what follows as constituting the cognitive component of the broader construct that psychologists refer to as *attitudes*.

Let us now go back to attitude-behavior consistency. Current research of the matter, as we saw above, aims at locating factors that can influence the extent to which our attitudes can reliably predict our behaviors. In this regard, one thing to note is that attitudes' predictive validity has been repeatedly found to be affected by measurement issues. The general observation is that, in order for people's attitudes to be useful predictors of their behaviors, there usually has to be a high level of correspondence, in terms of specificity, between the measure of the attitude and that of the behavior – a point that is currently referred to as the *principle of compatibility*.[13] Both attitudes and behaviors can indeed be described, and hence individuated for measurement

[11] Cf., e.g., Breckler (1984).

[12] Two seminal works on the function of attitudes, which inspired subsequent categorizations, are Smith et al. (1956), and Katz (1960).

[13] Cf., e.g., Fishbein & Ajzen (2010, ch. 2).

purposes, in more or less general terms. To exemplify, an environmental attitude can be described at varying levels of specificity as an attitude toward:

- Protecting the environment.
- Protecting the environment by purchasing an energy-efficient washing machine.
- Protecting the environment by purchasing an energy-efficient washing machine on Amazon.
- Protecting the environment by purchasing an energy-efficient washing machine on amazon over the next thirty days.

The principle of compatibility predicts that, in general, it is unlikely that an individual's intention to *protect the environment by purchasing an energy-efficient washing machine on Amazon over the next thirty days* – or her implementation of such intention[14] – will be reliably predicted by her general attitude toward *protecting the environment*. This does not mean, however, that general or global (as opposed to more specific) attitude measures are *never* useful in order to predict people's behaviors. Some evidence indeed suggests that global attitude measures can at times usefully predict aggregations of multiple behaviors that fall within the scope of the attitude or recurring behaviors, such as, e.g. recycling.[15]

Attitude-behavior consistency has also been found to be partly affected by various functional and structural properties of the attitude in question. As we saw above, the general function of attitudes is commonly thought to be that of fulfilling basic human motivations. Quite often, for instance, the holding of an attitude – say about a political candidate or a social practice – serves the purpose of expressing our values. In this regard, there is evidence that value-expressive attitudes are more likely to lead to behavior than attitudes held for other purposes.[16] As I also mentioned, attitudes are commonly taken to vary widely in their internal structure, as well as in their strength. In this regard, stability, accessibility, and internal consistency among their structural components (i.e. cognitive, affective, and conative) have all been found to increase attitude-behavior consistency by affecting the strength with which an attitude is held.[17] A further, and rather interesting, large set of findings concerns instead the way in which attitude-behavior consistency can be affected by either dispositional or situational individual differences. It has been shown, for instance, that our *knowledge* of the attitude in question proves relevant in this regard. The idea is that simply reflecting, via introspection, on the reason or reasons why we hold an attitude toward a given object can interfere with that attitude's ability to predict our behavior. Suppose that you happen to hold a positive attitude towards electric cars, but you do not quite know why you do, in the sense that you never really cared to inquire into this issue. If I now ask you to motivate your preference, you will likely begin to look for and come up with all kinds of reasons to support your attitude. Evidence however suggests that – even though, by so doing, your attitude toward electric cars will become temporarily more positive – your rationalizing activity will at the same

[14] The distinction between *behaviors* and *behavioral intentions* will be discussed below.

[15] Cf., e.g., Fishbein & Ajzen (1974), Weigel & Newman (1976), and Werner (1978). See, however Oskamp et *al.* (1991) for conflicting results.

[16] Cf., e.g., Fazio & Williams (1986) for the relation between political attitudes and voting behavior.

[17] For stability and accessibility, see the review of Cooke & Sheeran (2004); for internal consistency, see the review of Chaiken et *al.* (1995).

time artificially inflate it, thereby making it, in effect, *less* predictive of your future behaviors and intentions such as, e.g., your willingness to by an electric car.[18] A somewhat related, and equally puzzling effect can instead provide a vivid illustration of the way in which attitude-behavior consistency can be affected by situational, rather than dispositional, factors. The phenomenon in question involves *self-awareness* – i.e. the focusing of our attention on ourselves. The two most common methods for manipulating people's level of self-awareness are audiences and mirrors – i.e. subjects are asked to perform a task while either looking at themselves in a mirror, or standing in front of an audience. It has been found, however, that audiences tend to focus our attention on *public* aspects, whereas mirrors tend to focus it on *private* aspects of our self. To be relevant for present purposes is that attention to private aspects of the self has been observed to increase attitude-behavior consistency, whereas attention to public aspects thereof has been observed to increase the consistency of our behaviors not with our own attitudes, but rather with societal expectations – i.e. with our beliefs about the *norms* that regulate the behavior in question in our reference group. In other words, the attitudes that we report while looking at ourselves in a mirror are more predictive of our future behaviors than the attitudes we report while speaking in front of an audience.[19]

Over the decades, research has isolated a very large number of factors that are now known, or at least commonly believed by psychologists, to moderate the link between attitudes and behaviors. As it was the case for the persuasion data considered in sections 1.2 and 1.3, some researchers have tried to integrate this vast plurality of findings – which may this time be collectively referred to as the *consistency data* – into broader theoretical frameworks intended to explain them. The conceptual tools so far developed within such frameworks have proved very useful in order to design effective interventions aimed at facilitating behavioral change. There are currently two dominant theories of attitudes' influence on behavior: the *Reasoned Action approach* (henceforth, simply RA approach or model), and the MODE *model*.[20] In the remainder of this section, I will only sketch the outline of the former, and this is for two main reasons. On the one hand, the RA approach played a foundational role in setting up the conceptual framework later developed and fine-tuned by subsequent theories. The MODE model, in particular, was originally conceived as a way to address the limits of – and thereby serve as a critical counterpoint to – the approach out of which it developed – i.e. the RA model. The second, and more important, reason is instead directly linked to the specific purposes of this dissertation. Indeed, as we shall presently see, in predicting behaviors from attitudes, the RA model assigns a key role to the cognitive component of attitudes – i.e. beliefs.

The fundamental assumption of the RA approach is that, in general, the most direct predictor of a given behavior are our stated *intentions* toward that behavior, defined in terms of an individual's subjective probability of performing it.[21] This basic idea spawned more than a thousand studies, and underwent several modifications over the years.[22] Behavioral intentions, according to the RA framework, would in turn be predicted by only three basic factors – i.e. our *attitudes* toward the

[18] Cf. Wilson et *al.* (1989).

[19] Cf. Froming et al. (1982).

[20] Cf. Fishbein & Ajzen (2010) for the most comprehensive and up to date presentation of the RA approach. For the MODE model, cf., in particular, Fazio (1986, 1990), and Fazio & Towles-Schwen (1999).

[21] "The essential underlying dimension characterizing an intention is the person's estimate of the likelihood or perceived probability of performing a given behavior." (Fishbein & Ajzen 2010: 39).

[22] Cf., e.g., Fishbein & Ajzen (1975), Ajzen (1985), and Fishbein (2000).

behavior in question, *perceived norms*, and *perceived behavioral control*. The crucial aspect, for present purposes, is that each one of these three factors is taken by Fishbein and Ajzen to be based on three corresponding kinds of *beliefs*, as depicted in the Fig. 1:

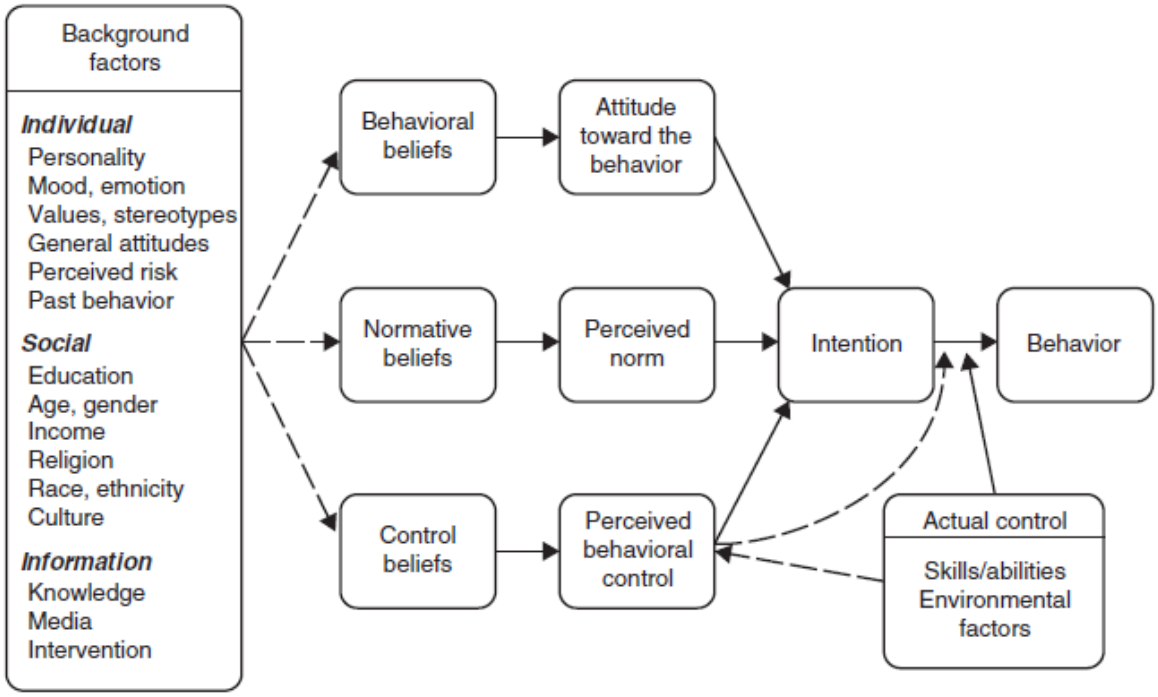

Fig. 1. The Reasoned Action approach. Source: Fishbein & Ajzen (2010: 22).

As one can see from this schematic presentation, the model posits that our attitudes toward a given behavior will be based on *behavioral beliefs* – i.e. our beliefs about both the consequences of implementing (or refraining from implementing) the behavior, and the value we attach to those consequences. Perceived norms are instead based on *normative beliefs* – i.e. our beliefs about whether relevant others would approve or disapprove of our performing the behavior (injunctive norms), or about what they would actually do in similar conditions (descriptive norms). Perceived behavioral control, finally, is based on *control beliefs* – i.e. our beliefs about our own ability to actually enact the behavior in question. To illustrate, let us go back to a previous example and suppose that John, a prospect customer, is currently considering buying a Tesla for Christmas from the local dealer. The RA framework suggests that the most reliable way to predict whether John will actually form an intention to buy the car is to gather information about John's relevant behavioral, normative, and control beliefs. Suppose, then, that John holds the following set of beliefs: B_1. Purchasing the electric car will likely make transportation more convenient for him. B_2. His family and friends will approve of his purchase. B_3. The local car dealer is currently selling the Tesla at an unbeatable price.[23] Under the above conditions, we have good reasons – indeed,

[23] This is of course and intentionally oversimplified illustration of a typically much more complex real-life situation. In particular, as shown in Figure 1, John's control beliefs will be affected by his *actual control* over forming the relevant intention, such as, e.g., his bank account not being currently empty.

according to the RA model, the best possible reasons – to predict that John will likely form the intention to buy the Tessla. From a strictly theoretical point of view, an appealing feature of the RA model is represented by its parsimony. On this model, as we just saw, John's behavior is a direct function of only three basic predictor variables. Indeed, while recognizing the importance of *background factors* – such as, e.g., past behavior, education, or knowledge – in explaining people's actions, the RA model holds that such factors will only influence our behavior indirectly – i.e. by affecting the way in which we form our beliefs about behavior, norms, and control.

3.2 The broad reach of behavior

As we saw in the last section, our behaviors are significantly influenced by our beliefs. The RA model has indeed proved very successful in many different behavioral domains. To date, over a thousand empirical studies conducted within this framework globally testify to its remarkable robustness in predicting people's behavioral intentions and actual behaviors.[24] In particular, evidence suggests that, just as posited by the model, behavioral intentions *are* indeed a valid proxy for behavior, and that it is therefore warranted to rely on findings concerning the former in order to predict the latter.[25] At least on this score, then, the lay intuitions we started from in section 3.1 proved largely correct – i.e. beliefs *do* shape behaviors, and can hence be relied upon in order to explain and predict them. Research has shown, however, that beliefs and behaviors are linked in reciprocal ways. So, the question arises: How about the other way around? The present section will focus on the opposite direction of the causal arrow. The general question on our agenda will hence be the following: do behaviors shape beliefs? As it is often the case in theoretical matters, a convenient entry point to this issue consists in raising a further question – i.e. where do beliefs come from? In this regard, the beginning of an answer is already provided by the schematic presentation of the RA approach considered above. Indeed, as shown in Figure 1, Fishbein and Ajzen explicitly acknowledge that *past behaviors* are among the background factors that affect the way in which we form beliefs. In other words, the specific ways in which we think and feel about various objects, according to the model, are not just a function of which other beliefs and attitudes we happen to hold; they are also, and importantly, affected by the actions and behaviors we happen to have performed. This hypothesis, as we shall see, is in line with a long tradition of research documenting the surprisingly vast array of behaviors that are now well known to influence our beliefs and attitudes.

Some of the actions that affect our beliefs and attitudes are performed non-consciously, often as an automatic response to basic physiological stimuli or bodily movements. Social psychologists refer to such responses as *implicit behaviors*. As I see a picture of a certain woman, I notice that my heart rate has accelerated, that I started smiling and automatically reached out to the picture to bring it closer and take a better look at it. I conclude that I must like the woman in question. The striking finding, in this regard, is that we are so used to the idea that our bodies provide us with

[24] Cf. Armitage & Conner (2001) for a meta-analytic review of 185 studies. Sparks & Shepherd (1992) is an application of the RA approach in the environmental domain.

[25] Sheeran (2002), a review of 10 meta-analyses, has found a strong mean correlation ($r = .53$) between behavioral intentions and actual behaviors. Experimental evidence indeed suggests that manipulating people's behavioral intentions or even merely asking people about them increases the likelihood of the corresponding behavior being enacted. Cf. Webb & Sheeran (2006), and Levav & Fitzsimons (2006), respectively.

reliable information about our attitudes, that I would have likely drown the same inference (I must like this women!) even if the physiological feedback in question had been bogus. In a seminal study, Valins (1966) provided male subjects with real time *false* feedback about their heart rate while watching 'sexually oriented stimuli' consisting in pictures of 'seminude females'. Needless to say, the pictures that subjects rated as more attractive were the ones associated with the false heart rate increase, and the ranking stayed the same when subjects were asked to re-rate the pictures several weeks later.[26] One thing to note about this study, however, is that Valin's subjects were consciously aware of the existence of a connection between their implicit behavior and the attitude object. So, the natural question is: would implicit behaviors still affect our attitudes toward a given object in case we were not even aware of the existence of a connection between the two? The answer seems to be positive. Strack and colleagues (1988) had subjects rate how funny a cartoon was while holding a pen in their mouth. In particular, participants in two different conditions were either asked to hold the pen between their lips without it touching their teeth, or to hold the pen between their teeth without it touching their lips. As a consequence, unbeknownst to the subjects, their faces assumed either a *frown* or a *smile* position – i.e. two bodily movements commonly perceived as respectively negative and positive. The results showed that, although the subjects did not consciously realize that they were either frowning or smiling while performing their task, the unwittingly frowning ones rated the cartoon more negatively, whereas the smiling ones rated it more positively.[27] Briñol and Petty (2003) reported similar findings with respect to students responses to persuasive messages. Apparently, being experimentally induced to either nod our heads up and down (a behavior typically associated with confidence) or shake them from side to side (a behavior typically associated with lack of confidence) while reading an argument affects whether we find the argument in question more or less convincing.[28] Whether or not we realize it, then, implicit behavior can significantly influence our beliefs and attitudes about various matters.

How do matters stand with respect to *explicit* behaviors – i.e. behaviors that we consciously and intentionally engage in? The origins of current research on this issue can be traced back to what is perhaps the most famous theory in contemporary social psychology – i.e. Leon Festinger's *Theory of Cognitive Dissonance*.[29] Festinger's fundamental idea was that we represent elements of our psychological lives in terms of cognitive elements he called *cognitions*. Cognitions are basically representations of how we feel, how we behave, and what we think about the environment (e.g. perceptions and beliefs). The crucial point is that such cognitions can at times be psychologically *inconsistent* with one another. This would be the case, for instance, if I were a committed environmentalist who prefers electric cars and believes that they are better for the environment, and yet ended up buying a diesel Ford instead of a Tesla. In a similar situation, according to Festinger, I would experience an aversive, uncomfortable state of arousal that he called *cognitive dissonance*. The basic tenet of his theory is that humans are naturally motivated to avoid or reduce the dissonance created by inconsistent cognitions. We can usefully construe this on a par with what happens in the case of other, more familiar, aversive states, such as e.g., hunger or thirst.

[26] Cf. Valins (1966). Taylor (1975) had instead women rate pictures of men after giving them false physiological feedback about galvanic skin response (GSR). Interestingly, this time the stimuli were not 'seminude' men but consisted in 'full-face color portrait photographs of clothed male graduate students'.

[27] Cf. Strack et *al.* (1988).

[28] Cf. Briñol & Petty (2003).

[29] Cf. Festinger (1957).

Just as we are naturally driven to reduce hunger by eating, and thirst by drinking, Festinger maintained, we are likewise driven to reduce dissonance by *changing* our cognitions in order to make them consistent with one another. In this regard, research has singled out several different situations that are likely to create dissonance. The most studied ones are *choice*, *effort justification*, and so-called *counter-attitudinal behavior*. Let us take a look at each one of them.

As much as we would often like not to, life requires us to make a myriad choices and decisions on a daily basis. Some are trivial – e.g. what to eat for dinner, what to wear at the party, or which movie to watch on Netflix – yet some are definitely more consequential – e.g. where and what to study, which car to drive, for whom to vote, or, for the most fearless amongst us, whom to marry. One thing that sucks about choosing (and sometimes, as in the case of some bachelors, prevents us from doing so) is that choices come at a nasty price which economists tellingly refer to as *opportunity cost*. If I buy the diesel Ford instead of the electric Tesla, I will enjoy the positive features of the Ford, yet at the same time I will have to relinquish the positive features of the Tesla – not to mention the fact that, since any car is likely to have at least *some* negative features, by purchasing the Ford I will also have to endure its negative features. This state of affairs is bound to create dissonance. In general, research has repeatedly shown that, as a consequence of my choice, my attitude toward the chosen item will become more positive, whereas my attitude toward the rejected one will become more negative – i.e. *after* my purchase, I will likely come to like the Ford more, and the Tesla less than I did *before* the purchase. This long-investigated phenomenon is known as the *spreading of alternatives*. It would appear that the very act of *choosing* something can set off a process that results in a change in our attitudes to bring them more in line with our past behaviors. Why does that happen? The standard answer is that, as past choices are not easily undone, changing our attitudes is by far the most expedient way to reduce dissonance. It must also be noted that the size of the choice effect has been found to be a function of both how important the choice in question is to us, and how hard it happens to be. In particular, the closer two options are in our pre-choice ranking, the more post-choice dissonance we will experience, and the more spreading of our preferences will ensue.[30]

As you may recall from section 1.1, according to Mercier and Sperber's interactionist view, human reason evolved primarily for social purposes. One of its main functions, in particular, would be to *justify* our beliefs and actions. Now, regardless of how one feels about their general proposal, it seems undeniable that a large portion of our mental resources is indeed devoted to motivating and defending our past decisions and actions. To be relevant for present purposes, however, is that we often engage in such activity in order to justify our actions not in the eyes of others, but to our own selves. A familiar situation, in this regard, is the one in which you worked very hard to obtain something that, once obtained, turns out to be way less cool than you expected – e.g. after months of courting, she finally agreed to a date, and you took her out to have dinner in the finest restaurant in town. Yet, alas, as the evening unfolds you slowly and painfully realize that her conversation bores you to death, as she appears to have no other interests other than stray cats. A gloomy thought suddenly dawns on you: "this was definitely *not* worth the hassle!" Unless you are a cat-loving bore yourself, this situation is likely to put you in a state of high dissonance, and the easiest way out of it, as we already know, is to change your mind about your date – e.g. "she is not that boring after all, she just had a very long day. Plus, I always kind of liked cats and,

[30] The choice effect was first demonstrated by Brehm (1956). Cf. Wicklund & Brehm (1976), and Cooper (2007) for reviews.

for a (short) while, even considered taking one in from the local shelter!" The moral here is that, as common wisdom suggests and research amply confirms, the effort we put into achieving a goal in life, can significantly affect our beliefs and attitudes toward that goal.[31]

Now consider one last familiar situation. Unless you are a very unusual human being, you must have found yourself in a circumstance or two in which you *said* something that you do not fully believe or ended up defending a position that you do not fully share. I bet, for instance, that in spite of your allergy (which, by the way, she knew nothing about), during that expensive dinner you dropped a few nice comments about how cute kittens are. In fact, chances are that, perhaps simply out of politeness, you even showed some sympathy toward a local politician's proposal to divest some money from your town's already crumbling sports facility and use it to build a new cat shelter, even though the truth of course is that, as a swimmer allergic to cats, you hardly feel that way. Social psychologists refer to public statements that directly contradict a belief or attitude that we currently hold as *attitude-discrepant* or *counter-attitudinal* behaviors. What you did at your romantic dinner hence counts as such and, as such, it is very likely to produce dissonance, and to therefore lead to attitude change as a way to reduce it ("let's face it, kittens *are* cute, and our town *does* need a new cat shelter!") The effect of counter-attitudinal statements on attitudes was first demonstrated in the late 1950s by Festinger and Carlsmith in a seminal experimental study that established what is now known as the *induced compliance* paradigm.[32] Participants in the study were initially asked to perform an extremely *boring* task consisting in rotating pegs on a pegboard according to a set of rules. After performing the task, they were induced by a clever cover story to *willingly* comply with a request of the experimenter – i.e. be hired to convince other (fake) fellow participants who had not yet performed the task, that the latter was indeed very *fun*. As I hope is clear by now, this was done in order to create dissonance. The experimenters indeed predicted that publicly stating that (contrary to what they thought) the task was very fun should cause some attitude change in that direction – i.e. after delivering the counter-attitudinal speech, participants should come to find the boring task slightly more enjoyable. So far so good. Yet now comes the interesting twist. According to Festinger's original theory, indeed, dissonance has a magnitude. In particular, the theory predicts that the more discrepant you perceive your behavior to be from your actual attitude, the more dissonance you will experience, and the more you will need to change your attitudes in order to restore consistency. To test this additional prediction, Festinger and Carlsmith created two groups. Whereas participants in the first group were offered only $1 to deliver the counter-attitudinal speech, participants in the second group were instead offered $20.[33] The experimenters this time predicted that while a low financial inducement ($1) should produce *more* dissonance and subsequent attitude change, a high financial inducement ($20) should instead produce *less* dissonance and less attitude change. Why? Well, their reasoning was that participants in the high financial inducement condition would have perceived their behavior as less discrepant with their actual attitudes than participants in the low financial inducement condition. After all, they had been well paid for the job of delivering the speech, and this 'cognition' – i.e. their being aware of this fact – should allow them to regard their behavior

[31] The effort justification effect was first demonstrated by Aronson & Mills (1959) with respect to attitudes towards groups, and was later documented in various other contexts, such as, e.g., our attitudes toward political issues (Wiklund et al. 1967), psychotherapy (Axsom 1989), and video games (Wan & Chiou 2010).

[32] Cf. Festinger & Carlsmith (1959).

[33] Let me remind you that the study was carried out in the late 1950s. Adjusting for inflation, the two financial inducements would today be approximately $ 10 and $ 200.

as driven more by the money than by their own true attitudes toward the task. As you may suspect, this is exactly what Festinger and Carlsmith found. Participants in the high financial inducement group changed their minds toward the task a lot less than did participants in the low financial inducement group.[34]

Festinger and Carlsmith's pioneering study had a disruptive effect on contemporary theorizing about behavior-attitude links. As we saw in section 1.2, the dominant approach to cognition at the time was learning theory. This general approach had no problem accommodating the idea that our mental lives seem powerfully affected by consistency considerations – i.e. that we often strive to achieve a balance between our attitudes and our behaviors. According to the received wisdom, however, it simply did not stand to reason that someone's attitudes could change more as a consequence of a small, as opposed to a large, financial incentive. To the extent that behavior, just as persuasion, was then commonly taken to be simply a matter of learning and reinforcement, behaviors induced by smaller incentives should 'obviously' lead to *less* attitude change, not more! Yet this is precisely what researchers working within the induced compliance paradigm kept observing, and this pretty much looked as a fatal blow to one of the most basic tenets of learning theory.[35] Moving on from historical considerations, however, the basic theoretical assumption that research on behavior-attitude links had confirmed up to that point was the following:

> Attitudes are affected by behavior and change to become consistent with it.

As it was to be expected, a new question then reared its head – i.e. is that unavoidable? Does inconsistency *always* lead to attitude change? It did not take long for researchers to realize that the answer was no – i.e. a mere logical inconsistency among our cognitions is not sufficient to produce dissonance and lead to attitude change. So – analogously to what happened in the case of attitudes' influence on behaviors – the new question became rather *when* do attitudes change following behavior? In this regard, subsequent research aimed at locating the limiting conditions under which behavior *does* influence attitudes soon hit on three crucial factors. One factor was found to be the individual's personal *commitment* to the counter-attitudinal behavior in question. If you write a counter-attitudinal essay while knowing that you will later have a chance to justify yourself by explaining that you did not really mean what you wrote, your behavior is unlikely to produce any attitude change – i.e. no commitment, no dissonance.[36] This finding is related to the second crucial factor. Indeed, your explanation of why you wrote the essay without really meaning what you wrote could of course appeal to the fact that you were forced to do it. If that were indeed the case, or – importantly for our purposes – even if you merely *perceived* it to be, then again, your behavior would not produce any attitude change – i.e. *no freedom*, no dissonance.[37] A final crucial factor that was soon found to affect attitudes' change following behavior is whether the counter-attitudinal behavior in question produces *unwanted* or *undesirable* consequences. If the next participant in line in the peg rotating task whom you had intentionally tried to dupe into

---

[34] Counter-attitudinal behavior has been observed to lead to attitude change in many real-life situations involving attitudes toward, e.g., free-speech (Linder et *al.* 1967) or local campus issues (Wakslak 2012).

[35] It is therefore unsurprising that – in spite of repeated replications of Festinger & Carlsmith's (1959) basic finding (cf. Cooper 2007 for an extensive review) – some authors attempted to question its validity by criticizing its underlying methodology. Cf. e.g., Janis & Gilmore (1965).

[36] Cf., e.g., Carlsmith et *al.* (1966).

[37] Cf., e.g., Linder et *al.* (1967).

believing that the task was a lot of fun had ostensibly not believed you, and failed to form a false belief because of you, you would not have experienced any dissonance and would not have come to like the boring task any better. Similarly, if you wrote a counter-attitudinal op-ed article on a newspaper while making clear that it was intended as a hoax, you may get bad press, but would not alter you attitudes as a consequence – i.e. nor harm, no dissonance.[38]

Over the years, research aimed at addressing the *when* question by carefully documenting the conditions under which behavior influences attitudes accumulated a wealth of findings that, in line with the organizational criterion so far adopted, we may collectively refer to as the *dissonance data*. As it was the case for the persuasion data already discussed in sections 1.2 and 1.3, and the consistency data considered in section 3.1, the *when* question eventually led to the *why* question, and researchers began to put forward different theoretical frameworks intended to make sense of the existing dissonance data, as well as to predict new ones. Although the ensuing debate over the complex cognitive mechanisms that underlie attitude change following behavior is still very much open today, it will be useful for present purposes to take a selective look at some of the main ideas introduced in the above-mentioned explanatory frameworks.

According to Cooper and Fazio (1984), the route to theoretical illumination consisted in taking a new look, or a new perspective on the dissonance data themselves. Why would a counter-attitudinal behavior affect our beliefs and attitudes only if it was freely chosen? And why did it need to bring about unwanted or aversive consequences? The key to solving the riddle, according to the two authors, had to be looked for in the notion of *responsibility* for the consequences of our own actions. This implied, in particular, that the consequence should not only be unwanted or aversive, but also, and crucially, *foreseeable* – a prediction that later research suggested to be on the right track.[39] Eddie and Cindy Harmon-Jones (2002) proposed an interesting *Action-Based Model* of dissonance, according to which the relevant attitude change would have both a *proximal* and a *distal* motivation. Whereas the proximal motivation – in line with Festinger's original insight – would be dissonance reduction, the distal motivation would be to prepare us for resolute action – i.e. to defuse the threat of inaction in front of a difficult choice. On their functional model, that is, cognitive dissonance would hence be at bottom an action-oriented state.[40]

Another influential family of frameworks has had the merit of shedding light on the important role played by the *self*-concept in the cognitive mechanisms responsible for dissonance. According to Aronson (1969), for instance, dissonance would only be aroused when you act in a way that you perceive as being at odds with what you happen to think of yourself – i.e. when your behavior, as it were, violates your expectations concerning the kind of person that you take yourself to be.[41] Most of us think of themselves as fundamentally good persons, and doing something bad, like throwing trash out of a car window on a highway put us in a state of dissonance because, or so we like to think, "that is not who I am!" It is worth noting, however, that anchoring dissonance to our self-esteem has a rather puzzling consequence, with far-reaching pedagogical implications. Some evidence indeed suggests that, in the sad case in which we unfortunately happened not to think much of ourselves, our own expectations will in effect be violated when we actually do something good! E.g. if a pupil is repeatedly told that she sucks

[38] Cf., e.g. Cooper & Worchel (1970).
[39] Cf. Cooper & Fazio (1984), and Goethals et *al.* (1979).
[40] Cf., e.g., Harmon-Jones & Harmon-Jones (2002) for supporting evidence.
[41] Cf., e.g., Aronson (1969).

at math, she will experience dissonance upon getting a good grade, and will end up being less confident about her manifested skill.[42]

A final framework definitely worth considering was put forward in the late 1960s by Daryl Bem.[43] What set his *Self-Perception Theory* noticeably apart from all other systematic attempts at explaining the extant findings – and makes it, to my eyes at least, one of the most fascinating chapters in the history of social psychology – was its radical questioning of Festinger's pivotal assumption that what drives attitude change following counter-attitudinal behavior would be a deep-seated need to reduce dissonance. Bem's general attitudes toward psychology had indeed been shaped by Skinner's behaviorism, and that is arguably what made his proposal an altogether different beast at the time.[44] In his view, there was no need to postulate a latent construct such as cognitive dissonance in order to account for what researchers had till then observed. The cornerstone of his proposal was indeed the basic observation that our mental lives are not completely transparent to us. Quite often, in particular, we seem to have very limited access to our mental states. As a consequence, he maintained, we frequently *infer* what our attitudes toward various objects must be based on the observation (or, as he put it, 'perception') of our own past behaviors and circumstances. It may well be the case, for instance, that you simply happen to have no clear-cut attitude towards a given environmental issue, say, air travel or water recycling. In similar situations, he claimed, if someone asks your opinion about the matter, your initial introspective search will not deliver any output, and you will hence automatically start looking for cues in your recollection of your own past actions and choices. The following two propositions, by Bem's own lights, constitute the heart of his theory:

> "Individuals come to "know" their own attitudes, emotions, and other internal states *partially* by *inferring* them from observations of their own overt behavior and/or the circumstances in which this behavior occurs. Thus, to the extent that internal cues are weak, ambiguous, or uninterpretable, the individual is functionally in the same position as an outside observer, an observer who must necessarily rely upon those same external cues to infer the individual's inner states"[45]

If this is on the right track, then there are plenty of situations in everyday life in which we have to become the observers of ourselves. Whenever this is the case, the processes by means of which we attribute mental states to *ourselves* will be the *same* ones that we normally rely on to attribute mental states to *others*. Although we will go back to this heretical idea, it is here worth noting that both Dissonance and Self-Perception make the exact same predictions in the induced compliance study. Theoretically, indeed, nothing prevents us from explaining the relevant attitude change by suggesting that, whereas subjects in the low financial inducement condition were just *interpreting* their past verbal behavior as evidence that their attitude toward the task must have been positive,

---

[42] Cf., e.g., Aronson & Carlsmith (1962). Other two frameworks that have so far emphasized the importance of the self-concept are Steele (1988), and Stone & Cooper (2001).

[43] Cf. Bem (1967, 1972).

[44] To my knowledge, Bem first made his ideas public in 1967, the same year in which Ulrich Neisser published his *Cognitive Psychology*, unanimously taken by scholars to represent – together with Noam Chomsky's (1959) review of Skinner's *Verbal Behavior* – the tombstone of behaviorism. To quote the great Bob Dylan, Bem must not have needed a weatherman to know which way the wind was blowing, and this, in my opinion, testifies to the admirable independence of his own thinking about psychological matters.

[45] Bem (1972: 2, emphases added).

subjects in the high inducement condition were instead discounting the same piece of evidence as not reflecting their true attitudes. Be that as it may with respect to that study, let me close this section by simply taking note of a fact that will become relevant in what follows: a long tradition of research in social psychology seems to have firmly established the truth of the following proposition:

Behaviors *shape* beliefs, and can hence be relied on to *explain* and *predict* them.

3.3 Opening the black box

A much-felt need in current research on nudging is that of reaching a better grasp on the various psychological processes that underly nudging interventions, and can hence be credited for making a difference to the behavioral variable of interest by bringing about the desired behavior change. As we have seen in section 1.4 above, the basic findings that nudgers today capitalize on can be traced to research on human behavior and decision making carried out by economists and psychologists starting from the 1950s. As we have also seen, although the great Herber Simon had long been advocating for the urgent theoretical need to bring a host of cognitive insights to bear more closely on the study of human economic behavior, psychological research only began to establish a systematic relationship with policymaking in the early 2000s, when – largely due to the favorable reception of Thaler and Sunstein's book by the large public – the idea that behavioral science might fruitfully be relied upon in order to inform the design of public policies began to spread on a large scale. The basic idea of the nudge approach, we said, is that by leveraging our steadily growing knowledge of humans' psychological makeup – i.e. of the kind of decisional procedures that we *actually* adopt (such as heuristics) as opposed to the ones that we ideally *ought* to adopt according to the standard model of rational choice – one can often affect people's behaviors by means of subtle adjustments to their so-called choice architecture – i.e. the physical or digital environment in which they make their choices. What we now call *behavioral policy* is therefore "premised on the idea that interventions in public policy should be based on a psychologically realistic picture of *human* behavior and its *causes*".[46]

From an historical point of view, then, nudging seems accurately described as the strongly desired offspring of a now long celebrated wedding between economics and psychology. According to some, however, this initially happy marriage has been going through a rough patch lately, and – as it is often sadly the case in such complicated marital matters – its offspring is currently being negatively affected by this suboptimal state of affairs. Elaborating on the present metaphor, here is how some authors have recently summarized the sorry unfolding of the events.[47] Soon after the wedding, or so they claim, economics began to dangerously neglect psychology. To be sure, he would keep publicly claiming to constantly rely on her valuable help, yet that was clearly just the 'official story'.[48] What really went down, they suggest, can be put as follows. For a short while,

[46] Marchionni & Reijula (2019: 56, emphases added).

[47] Cf., e.g., Berg & Gigerenzer (2010), Grüne-Yanoff (2016), Hansen (2019), and Marchionni & Reijula (2019).

[48] For some reason, it comes natural to me to picture *economics* as a male, and *psychology* as a female. I am of course aware that, especially at these times of widespread gender-related fanaticism, some might object to this idiosyncratic terminological choice. Yet this is *my* story, so you will have to bear with it.

the normal division of intellectual labor seemed to run smoothly. Psychology would provide economics with a wealth of information concerning which kinds of behavioral effects were most likely to be observed under which conditions, and economics would then rely on this valuable knowledge to make suggestions to policymakers about which types of policies would have the best chances of affecting a given behavioral variable of interest – say, a specific health, financial or environmental behavior. At a certain point, however, psychology became – at least in the eyes of economics – a somewhat overdemanding partner. All of a sudden, it was no longer enough for her to keep churning out empirical data – i.e. more information about which behavioral effects were most likely to be observed under which conditions. She now demanded to know *why* exactly *those* behavioral effects were most likely to be observed under *those* conditions. Economics at first took this to be just another of his partner's short-lived fads, raised his shoulders, and went on with business as usual. After all, policies produced by their partnership were often making a surprising and welcome difference to people's lives – what more could psychology possibly ask for? As we all know, however, marriage is about compromise, so economics tried to find a middle ground, and, mostly to please his quirky wife, made a habit of padding out reports of the newest nudging interventions with vague and handwavy hints toward possible cognitive explanations of the observed, post-nudge behaviors.

Stepping out of our metaphor, many today feel that many nudge-based approaches to behavioral change face a serious replicability problem, and that such state of affairs reflects negatively on nudging's status of an 'evidence-based policy'. What one indeed often observes is that a specific nudge which proved remarkably effective in one particular context, proves largely ineffective (or worst, backfires) in other, different contexts.[49] To illustrate, let us briefly go back to Schubert's (2017) taxonomy of green nudges, discussed in section 1.5 above. Type 2 nudges, as you will recall, are the ones that capitalize on the well-known power of *social norms* to change behaviors by leveraging our natural inclination to "follow the crowd" or imitate our peers in order to conform to various social expectations. As noted on that occasion, to date some of the most researched green nudges in this family make use of *descriptive norms* – i.e. typically, messages conveying information about which behaviors happen to be prevalent within a given reference group – to gently steer people toward the adoption of wiser environmental choices. Goldstein and his colleagues (2008), in particular, implemented a clever green nudge specifically designed to encourage indirect water conservation by avoiding unnecessary laundering. Their goal was to induce hotel guests to reuse their towels by exposing them to signs, placed in bathrooms, appealing to descriptive norms. The main finding of their two now famous experiments, as you might recall, clearly suggested that descriptive norm messages are more effective than simple messages appealing to environmental concerns – i.e. that providing hotel guests with information about whether (and to what extent) people in the same hotel (or the same room) reuse their towels (e.g. "almost 75% of guests reuse their towels") makes a larger difference with respect to their towel reuse rate than exposing them to messages appealing to environmental concerns (e.g. "help save the environment by reusing your towels").

Suppose now that you are a local and open-minded policymaker operating within a water stressed region of a Mediterranean, industrialized country – e.g. Spain. In view of the next cyclical summer draught, you are currently considering the introduction of new measures aimed at fostering water conservation. So, you casually start browsing through the web, consult your loyal pocket-sized

[49] Cf., e.g., Sunstein (2017), and Osman et *al.* (2020).

GPT assistant, and soon get wind of the existence of efficacious behavioral policies aimed at promoting water conservation by *indirect* means – such as, indeed, by increasing towel reuse rates in hotels. As the financial standing of your region happens to come mostly from summer tourism, and given that, in virtue of your position, you obviously have excellent connections with most managers of the countless hotels present on your territory, you immediately ask your secretary to schedule an appointment with the potential stakeholders. Already at the first meeting, however, the majority of your hotel manager friends, while admittedly intrigued by your 'indirect water conservation' idea, raise the obvious question: How do we pull that off? I.e. How do we gently induce our (often upscale) guests to reuse their towels? In spite of your usual poker face, the truth is that you have no idea. Yet, due to your line of work, you are also well aware that in business, just as in politics, answers are wanted fast. So, this is when the usually precarious yet diligent staff on your team enters the scene, runs a slightly more systematic search, and, given the time pressure, only comes up with Goldstein and colleagues' (2008) study. Experts confirm that the study has indeed been rigorously conducted, and by the end of the afternoon you happily phone your friends in management informing them that you finally have a 'scientific' answer to their sensible question – i.e. we can most effectively induce hotel guests to reuse their towels by appealing to descriptive norms. Although it takes a bit of explaining – especially to clarify to them what is meant by *descriptive* norms – they seem satisfied with your answer, and are now willing to join your venture. That very same night, however, a particularly overzealous member of your staff, perhaps lured by the tenuous prospects of a permanent job, decides to pass on the usual post-work pint with her colleagues to stay at the office a bit longer in order to learn more about indirect water conservation … And this is where things start going west (though in fact, as we shall see, east) for your project. The thing is that, by inquiring further in to the issue, your staff member happens to come across a *different* study, this time by two German researchers named Bohner and Schlüter. The rub is that, in 2014, Bohner and Schlüter reported running an equally rigorous replication of Goldstein and colleagues' (2008) original study, yet unfortunately came up with opposite results – i.e. their main finding is indeed that messages appealing to descriptive norms are *not* at all more effective than messages appealing to general environmental concerns. In fact, their field experiments show that the former are sometimes (i.e. in one study) even *less* effective than the latter at increasing towel reuse rates in hotel guests.[50] Importantly, the authors disclose that their replication was itself motivated by "inconsistencies across studies using similar designs" as the one adopted by Goldstein and colleagues.[51] The next morning, while informing you of her game-changing late-night find, your staff member diligently draws your attention to the fact that whereas Goldstein and colleagues' field experiment was conducted in a well-known national hotel chain in the U.S., Bohner and Schlüter's was instead conducted in a four-star hotel located in the center of a mid-sized town in the Northwest of Germany (study 1), and in a three-star hotel located in the outskirts of the same town (study 2). The thorny issue on your table is now the following: which policy should you go for? I.e. which of the two options is now more likely (than the other) to be effective at increasing towel reuse rates – appeals to descriptive norms or appeals to environmental concern? Based on the solid, yet limited evidence on your desk, it would seem that the answer depends on *where* you happened to find yourself at – i.e. if you were

[50] Cf. Bohner & Schlüter (2014).

[51] Othe replications of Goldstein et *al.* (2008) were conducted in Austria and Switzerland. Cf. Bohner & Schlüter (2014: 2).

in the *U.S.*, then you should probably go for *descriptive norms*, but if you were in *Germany*, it would probably be wiser to bet on *environmental concern* instead. The trouble, alas, is that you are neither in the U.S., nor in Germany – you hotels are in Spain! Sure, you could try to alleviate the despair caused by the sudden collapse of your confidence by telling yourself that Germany is at least on your same side of the pond (Europe), and that this might arguably make it slightly less risky to rely on Bohner and Schlüter's evidence by going for environmental concern. Chances are, however, that in spite of your dissonance-reducing psychological maneuvering, you might still end up feeling that, epistemically speaking, you are (almost) back at square one.

For obvious reasons, the above example – while of course based on real data – was simply meant as an impressionistic and admittedly oversimplified illustration of typically much more complex real-world situations. Hopefully, however, it should be enough to bring home the relevant point. The point in question is that, as Cartwright and Hardie (2012) put it, when it comes to assessing the effectiveness of nudges, the journey from "it worked there" to "it will work here" is far from a straightforward one, and reaching a final verdict about whether a given nudge will be effective *in our case* is a surprisingly complex and difficult task.[52] This is due to the fact that, just as policies in general, behavioral policies such ad nudges are highly *context dependent* – i.e. the evidence produced by their implementation is always tied to the specific environment in which they have been first introduced. An immediate epistemic consequence of this fact, as the above example illustrates, is that knowing that a given nudge produced an effect in the so-called *source* population within which it was initially implemented, although of course important, does not however straightforwardly justify us in casually assuming that it will hence produce a similar effect in the *target* population within which we are considering applying it. Another way to put this would be to say that nudges are extremely *implementation-sensitive* devices. In this regard, a distinction often appealed to in the evidence-based policy literature is the one between the *efficacy* and the *effectiveness* of a given policy. Whereas 'efficacy' is usually understood as "the ability of a treatment to produce benefit if applied ideally" (where 'ideally' means, roughly, in a controlled experimental setting), 'effectiveness' refers instead to "the benefit that actually occurs when a treatment is used in practice."[53] As this terminology can do some useful explanatory work for us, allow me to slightly tweak the standard meaning of the distinction, in order to more readily apply it to our present case. Let us then agree on calling *efficacious* a nudge that 'worked somewhere', and *effective* a nudge that can reasonably be expected to work – i.e. make a difference to our variable of interest (e.g. recycling) – in the specific environment in which we intend to implement it. Now, with respect to our previous example, finding out about Goldstein and colleagues' field experiment would clearly provide a policymaker with evidence of the fact that descriptive norms have proven an *efficacious* way of increasing towel reuse rates in hotel guests. Based on what we said, however, it should not make her particularly confident that the same nudge (descriptive norms) will be an *effective* way to achieve the relevant behavioral outcome in her specific situation, in that there seems to be a legitimate question as to whether the fact that a given nudge has worked in some context is really sufficient to justify its implementation in a different one. It follows from the above that, even in cases in which she already possesses some evidence that a given nudge has indeed proved efficacious, a crucial challenge that any choice architect has to face in

[52] Cartwright & Hardie (2012: 14). Cartwright and Hardie's observation applies to policies in general, yet nudges are here been considered as policies, and it therefore seems legitimate to take it to apply to them as well.

[53] Andrews (1999: 317), cited in Cartwright (2009: 187-88).

implementing that nudge will consist in finding ways to bridge (or at any rate reduce) the gap between its *efficacy*, one the one hand, and its *effectiveness*, on the other.

As I put it at the beginning of this section, many today feel that one can valuably contribute to bridge this gap by trying to reach a deeper understanding of the various psychological processes that underly nudging interventions. One important limit of most nudging studies, according to this line of reasoning, is currently represented by an almost exclusive focus on the *effects* of various nudging interventions, and a concurrent lack of effort devoted to the investigation of the specific cognitive *mechanisms* responsible for bringing about the observed effects. As the point has perhaps been most forcefully made by Grüne-Yanoff (2016), reconstructing parts of his criticism will help us bring it into sharp focus. Grüne-Yanoff's main worry is that while the enthusiastic proponents of nudge-based approaches currently present their recommended policies as 'evidence-based', they typically do not do much, or at least not enough, in the way of explaining through which mechanisms such policies operate.[54] Here is how he sees the situation:

> "The evidence cited for the justification of such policies tends to be of a particular kind: it shows that the policy intervention, in a particular environment, makes a difference to a behavioral variable of interest. Such evidence is either produced in controlled experiments, where the average effect conditional on the intervention is compared with the average effect conditional on non-intervention; or alternatively through observational studies, where the average effect is estimated through statistical analysis … *What is typically missing is any evidence about the underlying mechanisms* through which these policies affect behavior."[55]

He partly qualifies this general discontent by acknowledging that nudge supporters do at times consider, although in passing, possible explanations for nudged behaviors.[56] Yet he laments that, even when they do, they typically fail to provide any evidence bearing on the crucial issue of how the cognitive mechanisms appealed to in such explanations are distributed in the population targeted by the nudge. The overall idea, then, would be that although nudging practitioners are, perhaps understandably, very keen on reporting *that* a given nudge has worked in a given context, they hardly ever feel compelled to seriously venture into the kind of research that would indeed be required in order to clarify *why* the nudge in question has worked in that context. And yet, in his view, pending a sufficient amount of information about the cognitive processes that underlie the past efficacy of a given nudge, it is often not possible to predict its effectiveness in an intended target environment.[57] Although we are here focusing on Grüne-Yanoff's own way of articulating this worry, other authors have recently expressed similar views. Marchionni and Reijula (2019), for instance, have couched the same issue in the following terms:

> "Extrapolating behavioral policies requires evidence about the invariance of the causal relationship between the policy lever and the outcome. Knowledge of whether and when a

[54] Cf., e.g., Grüne-Yanoff (2016: 1).

[55] Grüne-Yanoff (2016: 1-2, emphases added).

[56] Although Grüne-Yanoff does not mention it, one of the examples he has in mind might arguably be Sunstein's (2016a) discussion of possible ways to explain the effectiveness of *green defaults*, to which we will soon go back. Cf. Sunstein (2016a, ch. 7).

[57] The same point, according to Grüne-Yanoff, applies to other properties of a given nudge – such as, e.g., its robustness, persistence, and welfare enhancing. For reasons of space, I will here only focus on effectiveness.

> given causal relationship remains stable typically … relies on evidence about the variables that mediate and those that modulate it, that is, on *mechanistic* evidence."[58]

The general emphasis, then, is on mechanisms. The next task on our agenda should hence ideally be to address a million-dollar issue – i.e. what is a *mechanism*? As it is well-known, however, there is currently a very complex debate going on in the philosophy of science on precisely this matter, and the attempt at sketching even the beginning of an answer to this difficult question falls way beyond the much humbler scope of the present dissertation.[59] For present purposes, it will hence be more advisable to scale down our ambitions by focusing instead on what is usually meant by this term within the current nudging literature and, in particular, on how the term is understood by the author whose views we are here considering. In this regard, Grüne-Yanoff suggests that we adopt the following minimal characterization, which he takes to do a good job at qualitatively capturing the key elements shared by most accounts:

> "A mechanism of a phenomenon consists of entities and activities, which are related in some organized fashion, and which are responsible for the production of that phenomenon."[60]

He illustrates his position by considering, among others, what in Schubert's (2017) taxonomy feature as Type 3 nudges – i.e. *defaults*. As you may recall from chapter 2, this family of nudges relies on the well-documented existence of a *default effect* – i.e. people's observed strong tendency not to change an option selected by someone else, even in cases where the cost of so doing would be very low. Based on what we said, the relevant question is: why do they? Why do people stick with defaults? In this regard (as we also noted in chapter 2) there are currently several theoretical explanations of the default effect on the market, each one pointing to a different psychological mechanism. Grüne-Yanoff focuses on two in particular: the one appealing to *inertia* (the so-called *status quo* bias), and the one relying on *implicit recommendation* (extensively discussed in section 2.5 above). His main point, as applied to the present case, is that our ability to predict the future effectiveness of a given default will crucially depend on our possession of sufficient evidence about which mechanism was actually responsible for bringing about the observed behavior in its past implementations. This is due to the fact that different mechanisms will generally require different *background factors* to be in place in order to produce their effects.[61] To see how this works, let us briefly consider the two above-mentioned explanations. According to the first one, as active choices call for a careful evaluation of the available options (together with ways to compare them), people would stick with the default in order to reduce the cognitive effort involved in so

---

[58] Marchionni & Reijula (2019: 62, emphasis added).

[59] Accessible introductions to this notion are Illari & Williamson (2012), Andersen (2014), and Craver et *al.* (2024).

[60] Grüne-Yanoff (2016: 466-7). The characterization is adapted from a similar one extensively discussed in Illari & Williamson (2012). It is worth noting that, in adopting it, he further points out that "when I speak of mechanism … I mean *models of a mechanism*, which typically idealize some aspects and are abstract representations of a mechanism." *Ibid.* (467). Again, philosophers of science have long grappled with the many theoretical issues raised by models' ontological and epistemological status. Two accessible introductions to this complex debate are Craver (2006), and Frigg & Hartmann (2020).

[61] What Grüne-Yanoff refers to as *background factors* are of course the variables that mediate and modulate the causal relationship between a nudge (the 'policy lever') and its outcome mentioned by Marchionni and Reijula (2019) in the passage quoted above.

doing. This mechanism arguably requires, as a relevant background factor, that the nudgee be actually *uncertain* about her preferences, and hence in need to *make* such an effort in order to reach her final decision. Under this explanation, then, someone whose ideas on the matter are already well-formed and stable will be unlikely to go for the default option.[62] According to the second explanation, on the other hand, people would stick with the default option not because of the deliberation costs involved in choosing, but mostly because they tend to interpret that option as an implicit recommendation or endorsement coming from the choice architect. As it has been repeatedly noted[63], this mechanism arguably requires, as a relevant background factor, that the nudgee actually possess a sufficient level of *trust* in the choice architect. It follows that, under this explanation, someone who has reasons to distrust the nudger will be unlikely to rely on the default option.[64] Suppose now that a given default worked in a given context, and we are currently trying to predict whether it will be effective in a different context within which we are now considering implementing it. The kind of evidence we need, according to Grüne-Yanoff, should tell us which one of the two mechanisms best accounts for its past effectiveness. Intuitively, if the mechanism in question is *inertia*, we should expect that low levels of *preference uncertainty* will negatively affect its operation; whereas if mechanism in question is *implicit recommendation*, we should expect that its operation will be likewise negatively affected by low levels of *trust*. Our task, at this point, will hence consist in finding out how matters stand with respect to these two background factors (preference uncertainty and trust) in the context within which we are considering implementing our default.

As it was the case with our earlier example involving descriptive norms, the one sketched above is an intentionally oversimplified picture – a toy model, if you will – of a typically very complex real-world scenario. In particular, it is of course always possible that *different* mechanisms will in fact operate in tandem within the *same* population or individual – and when this is the case, we would of course ideally need to find out exactly *which* mechanisms are simultaneously operating, as well as which one of them, if any, appears to have a stronger impact on behavior. In spite of such simplifying assumptions, however, the above toy model should be sufficient to illustrate the general line of reasoning exemplified by Grüne-Yanoff's criticism. The fundamental idea, in a nutshell, is that a promising way to shorten what Cartwright and Hardie (2012) refer to as the long journey from "it worked there" to "it will work here", according to many, is to systematically rely on psychology to address head-on the question of "*why* it worked there" – i.e. to open nudges' black box, and take as good a peek into it as we possibly can. In order to do this, we will need psychological research to provide us with evidence about the cognitive mechanisms that underlie nudges' effects on people's behavior. The final, and perhaps easily predictable, question, at this point, will be as follows: *how much* mechanistic evidence do we need in order for our nudging interventions to rightfully count as legitimate instances of evidence-based policy? Although this issue is bound, by its very nature, to remain open, we can close the present section by pointing to the *sufficiency principle* that Grüne-Yanoff himself suggests that we adopt on the matter:

> "A policy is based on sufficient mechanistic evidence if it takes all available mechanistic evidence into account, where availability is constrained by current theoretical and

[62] Cf., e.g., Bronchetti et *al.* (2013) for an example of this situation in the case of saving behavior.
[63] Cf, e.g., Sunstein & Reisch (2014: 141), Schubert (2017: 336), and Congiu & Moscati (2021: 205).
[64] Cf., e.g., Brown & Krishna. (2004) for an example of this situation in the case of consumer behavior.

technological feasibility. If information of this sort does not enter the discussion at all, these policies cannot and should not be described as 'evidence-based"'.[65]

3.4 Beliefs matter

According to the now familiar story retraced in section 1.4, by the early 2000s most natural born influencers operating both in the public, and the private sector had enthusiastically acknowledged the existence of a brand-new policy game in town – i.e. *nudging*. The tools of their ancestral trade had grown to include what appeared to be a very promising way to more effectively exercise an ability they had mastered since the down of time: intentionally alter people's behavior without resorting to coercion; making it more likely that they will think and act in a predictable way. In many ways, the freshly unearthed compliance-gaining technique had indeed proved surprisingly more powerful than good old persuasion. Intervening on people's beliefs and attitudes, in the eyes of many, seemed no longer necessary in order to affect their choices – at least in the short run. As someone put it, however, "in the long run we are all dead", so most governments and private companies around the world embraced the 'gentle push' approach to behavior change, and actively began to systematically nudge their citizens or clients in various directions; for good, and sometimes for bad. A couple of decades have gone by since then, and we seem to have now entered a second phase of more thoughtful reassessment of nudging's power, dangers, and limits. Serious nudging research, however, while currently thriving, is still in its infancy. In particular, its highly interdisciplinary nature makes it the case that its subject matter – nudge-based policies – can be approached from many different, though not necessarily conflicting, disciplinary angles. In this regard, as I hope to have made clear in the last section, research on nudge-based policies has recently highlighted a pressing need to engage more closely with the *psychology of nudging* or, as some have more explicitly suggested, to 'put the psychology back in nudging'.[66] The very same behavioral outcome, as we saw, will indeed typically admit of several possible explanations, each appealing to a different cognitive mechanism. Under such circumstances, figuring out which one of them was most likely responsible for bringing about (or failing to bring about) the relevant behavior is of crucial importance in order to be able to forecast whether a given nudge will be effective in a different context. As in the case of other kinds of policies, it has been noted, "when it comes to nudging, to know 'what works' you need to know *why* it works."[67] Another useful function that research on the psychology of nudging may valuably perform, as we saw in chapter 2, is to defuse, or help resolve, ethical debates. Some indeed worry, on various grounds, that nudges might end up clashing with widely held human values such as liberty, autonomy, respect, and dignity. Many of the moral objections so far pressed to nudge-based policies appear to be motivated by widely sharable concerns, and to therefore deserve a fair hearing. The psychology of nudging, however, clearly bears on such debates, as it seems that a fair assessment of the many ethical issues raised by the use of nudges in policy can only be reached through a preliminary solid understanding of the many different ways in which nudges can affect our everyday decisions and choices.

---

[65] Grüne-Yanoff (2016: 480).
[66] Marchiori et *al.* (2016).
[67] Hansen (2019: 11, emphasis added).

The good news, with respect to the above, is that one need not look very far in order to locate the building blocks out of which a mature psychology of nudging can be developed. This is due to the fact that, as we saw both in the first and in the present chapter, scientific psychologists have been interested in exploring ways to encourage behavior change since their discipline moved its first steps. Indeed, whereas David Halpern, the founder of the Behavioral Insights Team, reminds us that "people have been 'nudging' each other for as long as mankind existed"[68], others have gone as far as to explicitly deny that – even though various behavioral policies are currently being grouped under this catchy umbrella term – nudging itself can meaningfully be regarded as an entirely new research area. On this last point, the following two passages from Marchiori and colleagues (2016) are worth quoting in full:

> "Nudging elaborates … on previously existing knowledge about automatic psychological processes and related phenomena. As such, *nudging is not a new research field*, but a clever application of knowledge on behavior change and decision-making, that is now finding its way into policy making and consumer welfare."[69]

> "We posit that the debate on nudges is due to the lack of more in-depth research and knowledge surrounding several *unanswered questions* in nudging, which in part arose from its new application in fields such as policy making and consumer welfare. In other words, these legal, political, scientific and economic debates stem from the scarce work and insights into pressing issues surrounding the application of *well-established psychological knowledge* in policy making."[70]

Whereas we will take care of one of those 'unanswered questions' in due course, it must be noted at this stage that a very sizable portion of the 'well-established psychological knowledge' that Marchiori and colleagues are here referring to arguably comes from research conducted in social psychology over the last several decades, and, for this reason, closely considered in sections 3.1 and 3.2. Indeed, as we saw there, since its early days, social psychology has been deeply interested in experimentally investigating the many, and often counterintuitive, connections between our attitudes and behaviors. Due to those seminal explorations, in particular, we now possess solid reasons to believe that the attitudes we happen to hold, and the behaviors that we happen to engage in, are causally linked in reciprocal ways. This being said, the current state of knowledge on the complex attitude-behavior relationships that have been demonstrated to characterize our mental lives has steadily continued to evolve to this day. As a consequence, the present section – and in fact, retrospectively, this whole dissertation – is premised on the guiding idea that the long-standing tradition of research on attitude-behavior consistency, and the ongoing nudging research can helpfully illuminate each other. This thought can be unpacked into the following two claims. On the one hand, the psychology of nudging can and should be seen as the most recent chapter of a much larger story, to which it can therefore valuably contribute. On the other hand, as I will begin to show in this section, future research on the psychology of nudging can profit a great deal from building on previous results stemming from social psychology. In

[68] Cf. section 1.4 above.
[69] Marchiori et *al.* (2016: 3, emphasis added).
[70] *Ibid.* p. 4, emphasis added).

particular, I think that many of the ideas and findings discussed in the first part of this chapter – together with contributions to be discussed in the next section – can usefully be appealed to in order to land plausibility to the line of reasoning consisting in the following two theses: (1) To the extent that – as it has been amply shown – 'beliefs *shape* behaviors, and can hence be relied on to explain and predict them'[71], it seems only reasonable to expect that people's beliefs will have a sizeable impact on the nudging process. Moreover, and by parity of reasons, (2) to the extent that – as it has again been repeatedly shown – 'behaviors *shape* beliefs, and can hence be relied on to explain and predict them'[72], it seems just as reasonable, although perhaps less intuitive, to likewise expect that the nudging process will itself have a measurable impact on the formation of people's beliefs. Let us hence consider thesis (1), and its practical implications, in the remainder of this section, and move on to thesis (2) in the next one.

In section 1.4 above, as you will recall, I suggested the opportunity of drawing a distinction between two different routes to behavior change: an *internal* and an *external* one. The internal route, I said, is represented by persuasion, and it aims at altering people's behavior 'from the inside' – i.e. by intentionally intervening on the *beliefs* and *attitudes* that oftentimes lead people to form intentions and make corresponding decisions and choices. The external route, on the other hand, aims instead at achieving the same goal 'from the outside' – i.e. by intervening on the physical or digital *environments* in which those intentions, decisions and choices happen to be formed and made. According to this way of looking at things, as I noted on that occasion, nudges are most naturally regarded as constituting an external route to behavior change, as they do not produce their effects (when they do) by directly targeting people's beliefs and attitudes, but rather by carefully reshaping a specific aspect of their environment – i.e. its choice architecture. Although I still regard this distinction as theoretically useful, the above way of couching it may nonetheless, in the present context, run the risk of being somewhat misleading. In particular, as nudges intended target are *behaviors*, based on the above distinction someone may be tempted to assume that people's conscious (or unconscious) beliefs are therefore largely irrelevant to the nudging process. This inference, however, is far too hasty, and it should therefore be resisted. Indeed, already at an intuitive level, nudging does not of course take place in a void, and the nudgee herself is obviously anything but a blank slate. On the contrary, she will normally hold – or at any rate take herself to hold – lots of different beliefs; and such beliefs, common sense suggests, should be partly responsible for having shaped many of those very behaviors that choice architects often aim at changing by means of nudging. Moreover, the robust research tradition documented in section 3.1 above has long gotten us used to the idea that – although a host of (intensely investigated) moderating factors are now known to either increase or decrease attitude-behavior consistency – we are nonetheless justified in accepting the validity of some adequately qualified version of the so-called *principle of cognitive consistency* – according to which, as you might recall, people's actions and behaviors will normally be largely consistent with their stated attitudes and opinions. As a consequence, as I put it above, it would seem eminently reasonable to expect that even the behaviors targeted by nudges, although often automatic in nature, will nonetheless be affected by the beliefs and attitudes that the individual being nudged happens to hold. Is there any evidence that this is indeed the case? The answer is: yes, plenty.

[71] Cf. section 3.1 above.

[72] Cf. section 3.2 above.

Some of the existing evidence that what beliefs people happen to hold has indeed a measurable impact on the nudging process can conveniently be laid out by briefly demonstrating how such beliefs have been found to interfere with the effectiveness of each and every one of the three types of green nudges in Schubert's (2017) by now hopefully familiar taxonomy. Let us then start with Type 1 nudges – i.e. the ones that leverage people's natural desire to maintain a positive self-image. As we saw in section 1.5, a currently practiced way of doing this in the context of consumer behavior consists in raising the level of consumers' environmental awareness by increasing the salience of certain environmentally-relevant product features by means of *eco-labelling*. Indeed, as we noted, for us Humans – as opposed to Econs – consumer choices do not normally respond to financial considerations only. They are also an import way to express our opinions and values, and, crucially, to cultivate a positive self-image by acting in accordance with them. The rub of course is that, needless to say, we do not all happen to share the same opinions and values! As a consequence, how a given environmental message will resonate with different people will largely depend on how they think and feel about the environment. In other words, it is to be expected that relying on environmental concern to promote, say, energy-efficient technology will not always pay off, as some people, for whatever reasons, simply do not share such concerns. With respect to the kind of nudge that we are here considering, this implies that labelling a product as environmentally friendly can make it either attractive or unattractive to a consumer depending on what her prior beliefs and values happen to be. Gromet and her colleagues (2013), for instance, investigated the effects of political ideology on energy-related consumer choices in the U.S., a country where the public opinion has long been known to be heavily polarized on environmental issues such as global warming.[73] What one of their studies found is that, in a real-choice consumer context recreated in their lab, conservative leaning subjects were less likely to buy a more expensive and energy-efficient fluorescent light bulb, yet, crucially, only when the latter was labeled with an environmental message ("Protect the Environment"), as opposed to a condition in which the label only reported cost information.[74] Why did eco-labelling backfire in this case? The most likely explanation, as the authors themselves acknowledge, is that eco labels can at times actually push customers away, rather than attracting them, as they happen to convey ideas and values that some people simply do not wish to be associated with.

Let us now consider green nudges of Type 2 – i.e. the ones that leverage our inclination to "follow the crowd" by relying on messages appealing to (mostly descriptive) social norms. As you might recall from section 1.5, in the U.S. home energy reports (HERs) have proved promising for steering people towards wiser energy consumption by exploiting the well-known effects of *peer comparison* – i.e. by providing feedback which compares a given household's energy use to the one of neighbors with similar-sized homes and heat type. As it was the case with eco labels, however, even the effectiveness of this largely successful nudge has been found to be negatively affected by ideological priors. Costa and Kahn (2013), for instance, provided evidence that political ideology makes for a crucial mediating factor in HERs effectiveness. According to their results, energy reports are indeed not only two to four times more likely to have an impact on the energy-consumption patterns of political liberals, than on those of conservatives, but this latter segment of the population is also more likely than the former to opt out of receiving HERs, as well as to

[73] Cf., e.g., McCright & Dunlap (2011).
[74] Cf. Study 2 in Gromet et *al.* (2013).

report disliking them.[75] This last point seems especially relevant for our present purposes, as it highlights the fact that a given nudge's effect can be influenced by the particular way in which people feel about the policy in question (in this case, HERs). In this regard, a field experiment conducted in The Netherlands by Dewies and his colleagues (2021) tested nudges intended to increase low levels of compliance with a policy requiring employees of a local government department to wear an identifying lanyard with their badge for security reasons.[76] One of the implemented nudges relied on social norms. Although the study provided insufficient support for the effectiveness of this nudge, the instructive point for us is rather that a qualitative survey administered by the experimenters amongst the nudgees revealed that a large portion of them reported not to believe in the policy's effectiveness or necessity, and to find the nudge in question overly paternalistic, in that it apparently made them feel treated like kids. Dewies and colleagues' conclusion is aptly summarized by the title of their article: nudging is ineffective when attitudes are unsupportive. It will also be instructive to quote the following sentence from their discussion section: "We advise researchers as well as practitioners to survey and reflect on the target group's attitudes and preferences before the start of an experiment testing nudges."[77]

How about Type 3 green nudges – i.e. defaults? The issue of how people's prior beliefs affect defaults has already been partly discussed in the last section, while considering Grüne-Yanoff's (2016) criticism of the current state of the art in nudging research. On that occasion, we pointed out that different mechanistic explanations of defaults' inner workings will also require different background factors to be in place in order for the default to produce its effects. One such factor, as you will recall, were in effect people's beliefs about matters relevant to the default's domain. In particular, we said, someone whose ideas on such matters are already stable and well-formed will be unlikely to stick with the default option, under the cognitive-effort reduction mechanism. At least under that particular explanation, indeed, low levels of preference uncertainty should negatively affect the default's effectiveness. Although evidence for this claim has already been noted in that section[78], a couple of additional examples considered in Sunstein (2017) are here worth mentioning to buttress the point. The first one concerns marital names.[79] In the U.S., both men and women are defaulted into retaining their pre-marriage surnames. In spite of this, while the vast majority of men stay with the default option, about 80% of women[80] reject the default, and actively decide to change their name – i.e. for most women (in current U.S. culture) the default rule does not stick. Evidently, their preferences are already strong and clear, and they do not need to exercise much of a cognitive effort in order to ascertain them. Prevailing social norms, apparently, are in this case likely operating a sort of (stronger) counter-default. Beliefs about such norms, then, can often influence defaults' effectiveness.[81] The second example concerns French fries.[82] Just and Wansink (2009) examined how defaults would affect elementary school kids' food choices. Kids were served a lunch which included French fries as the default option, yet they were asked whether they wanted to swap them with apple fries. Needless to say, 95% of the

[75] Cf. Costa & Kahn (2013).
[76] Cf. Dewies et *al.* (2021).
[77] *Ibid.* p. 223.
[78] Cf. fn. 62 above.
[79] Cf. Emens (2007), discussed in Sunstein (2017: 8-9).
[80] This figure refers to college graduates.
[81] According to Fishbein & Ajzen's RA approach discussed in section 3.1 above, as you will recall, beliefs about norms (*normative beliefs*) are indeed a strong predictor of behavioral intentions.
[82] Cf. Just & Wansink (2009), mentioned in Sunstein (2017: 10).

kids stayed with the default. Two days later, the nudge was reversed – i.e. this time apple fries were the default option, and kids were asked whether they wanted to go for French fries instead. Even in this case, one arguably does not need a PhD in developmental psychology in order to confidently forecast the effect of this move: 96% of the kids opted out of the default option. Evidently, beliefs about which foods are more delicious, just as beliefs about which social norms are in place within our reference group, are often much stronger than defaults.

As I hope we can all agree on by now, nothing beats French fries. Jokes aside, I take the moral of the above examples to be that, as anticipated above, which beliefs people happen to hold constitutes a crucial mediating factor in the operation of many nudge-based policies. Indeed, as we just saw, people's extant beliefs – their *ideological priors* – appear to feature prominently amongst the primary causes of nudges ineffectiveness. An explicit recognition of this important fact has recently made its way into a theoretical article authored by Hauser and his colleagues.[83] Their main contention is that, in order to design effective nudges, one should take into account the targeted subjects' reported *beliefs* over and above their behaviors and the context within which they act – they accordingly refer to behavioral interventions that capitalize on this suggestion as *budges*. In advocating for their position, they introduce a *Beliefs-Barriers-Context* (BBC) *framework*, which they conceive of as a decision rule that both practitioners and academics can and should rely on to design more effective nudges, and which can be schematically represented as shown in Fig. 2.

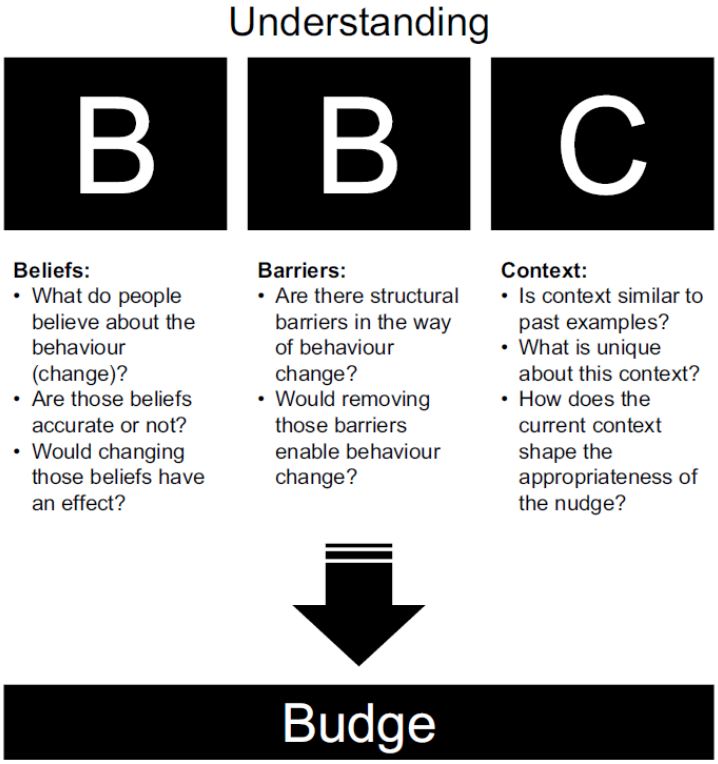

Fig. 2. The BBC framework. Source: Hauser et al. (2018:19).

[83] Cf. Hauser et *al.* (2018).

Hauser and colleagues think of their proposal as "a first step towards a comprehensive nudge theory."[84] However, insofar as both the importance of barriers to behavior change, as well as of the context in which a given behavior occurs have already been variously discussed within the present work[85], I will not further expand on these two aspects of their BBC framework. Indeed, whatever the other merits of Hauser and colleagues' recommended decision rules, to be relevant for our present purposes is rather their explicit recognition of the important fact that people's existing beliefs do constitute an essential parameter to be carefully taken into account when intending to design an effective nudge. The contents of such beliefs, as we saw above, can range over different subject matters, such as, e.g., the specific behavior targeted by the nudge, the nudge itself, or, as we shall see in the next section, one's own past actions. It is then time to move on and consider the second of the two theses which together constitute the theoretical backbone of the present dissertation.

3.5 Nudged beliefs and Self-Knowledge

The present dissertation is premised on the general idea that a long tradition of research in social psychology, and the promising yet still fledgling research on nudging can shed helpful light on each other. In particular, I suggested that the psychology of nudging, while constituting one of the most recent chapters in a much larger story – and indeed partly for this very reason – can profit immensely from building on previous, hardly-acquired knowledge of the many ways in which our attitudes and behaviors are causally linked in reciprocal ways. Such knowledge, as I tried to show in the last section, abundantly supports the idea that people's existing beliefs should have a sizeable impact on the nudging process – a hypothesis that current research on nudge-based policies, as we saw, has so far largely confirmed. That same knowledge, however, as I will suggest in the present section, lends support to a further, and at present still largely unexplored hypothesis. In order to set the stage for the latter, let us briefly go back to a forward-looking contribution mentioned in the last section. In their theoretical article, Marchiori and colleagues (2016) highlight four pressing questions about the core underlying mechanisms of nudging, which they regard as standing in urgent need of further investigation. The four 'yet unresolved questions', according to their agenda, concern the following issues: (1) the role of transparency, (2) the definition of the choice set, (3) freedom of choice, and (4) the consequences of nudging.[86] In line with Grüne-Yanoff's (2016) criticism considered in section 3.3 above, Marchiori and colleagues advocate for the need of more in-depth research on these issues, and they hold, as I do, that valuable insights towards resolving them should be looked for in current and past psychological investigations. Now, the new hypothesis to which I was just referring pertains to the fourth item on their list – i.e. the *consequences* of nudging. Here is what they have to say on the matter:

[84] *Ibid.* p. 20.
[85] Cf., e.g., sections 1.4 and 1.5 above.
[86] Cf. Marchiori et al. (2016: 4-5). The items have here been differently ordered for expository reasons.

> "While nudging interventions have been shown to have the potential to modify behavior in a specific decision-making situation … a next step in nudging research is to focus on examining long-term effects of nudging. No research so far has investigated the *consequences* of nudging interventions *from a psychological perspective*."[87]

This is precisely what I intend to do. In particular, what I will argue is that, based on our current knowledge of attitude-behavior links, we should expect one of those consequences to consist in a more or less sizeable shift in the nudgees' beliefs, over and above their behaviors. In the last section, as you will recall, I put this hypothesis as follows. To the extent that – as it has been repeatedly shown – behaviors *shape* beliefs, and can hence be relied on to explain and predict them, it seems reasonable to suppose that the nudging process will have a measurable impact on the formation of people's beliefs. So, let us take a closer look at this idea.

According to one of the godfathers of current social psychology, Richard Nisbett, over its long past, but especially over its short history, psychological research has produced three major insights into the way in which our minds work. Although I promise to disclose the second one in due course, the one that he evidently finds appropriate to mention first reads as follows:

> "Our understanding of the world is always a matter of *construal* – of *inference* and *interpretation*. Our judgments about people and situations, and even our perceptions of the physical world, rely on stored knowledge and hidden mental processes and are never a direct readout of reality."[88]

For present purposes, forget about perceptions of the inanimate physical world, and focus instead on our judgments about those funny animate objects that make our lives worth living – i.e. *people*. The place to start from is the simple observation that, at least in most everyday circumstances, such judgments *do* indeed feel like 'direct readouts' of reality. As you spot Clara amongst a crowd of youngsters walking towards your car after attending her best friend's birthday party, you are certainly not aware of any tacit inference or 'hidden mental process' leading you to the conclusion that your daughter is sad. You just seem to *see* that she is. Indeed, for all you can tell, this thought just popped up in your mind as you were consciously mulling over boring, work-related stuff. And yet – Nisbett tells us – in spite of its hardly questionable phenomenology, several decades of experimental research on human social cognition strongly suggest that your apparently snap judgment about Clara's current mood *is* indeed very likely the (conscious) output of a possibly rather long chain of (normally unconscious) inferential steps, each one plausibly taking as input a vast array of subtle cues to which you were not intentionally paying the slightest bit of attention. Suppose now that you side with Nisbett (and the vast majority of contemporary psychologists) on this score – i.e. suppose you accept that, although it certainly does not feel that way, most of your 'spontaneous' everyday judgments about people are not in fact a 'direct readout' of anything, but rather a matter of inference and interpretation. "This sounds plausible", you may think. "Yet, *surely*", you are even more likely to add, "*I* am not one of those people!" In other words, you are

---

[87] *Ibid.* p. 7., emphasis added.

[88] Nisbett (2015: 15, second and third emphases added). The *third* insight, in case you were wondering, is the following: "Many of the most important influences on our perceptions and behavior are hidden from us. And we are *never* directly aware of the mental processes that produce our perceptions, beliefs, and behavior". *Ibid.*

probably willing to bet your life on the fact that, 'surely', your own judgments about whether *you* are sad (or happy, or worried, or in love) are not at all the result of *inferring* or *interpreting* anything. "Look, when I am in love, I can't be wrong – end of story!" Although I usually try to steer clear of what a wise philosopher once mockingly dubbed the 'surely operator'[89], I have to admit that, in the present case, I partly sympathize with your frustration. The rub, alas, is that my argument crucially depends on the claim that – albeit with an important proviso to be presently noted – on many everyday occasions, you are *exactly* (the functional equivalent of) one of those people! The proviso in question is that, in spite of the above example, the argument does not require this claim to apply to our affective states, such as emotions, feelings or moods, but only to (part of) a limited subset of our cognitive states – i.e. *beliefs*. So, bear with me a bit longer, and yet, please, in considering the positions that I will introduce, try to beware of the 'surely operator'!
Consider the following remark on nudges, coming from Oliver's (2017) extensive reconstruction of the economic and psychological origins of behavioral policy:

> "Nudges are intended … to work with automatic decision-making processes. That is to say, they are not *intended* to change people's opinions about their behaviors in an overt, explicit way through information or persuasion campaigns and education programs, such that people come to a deliberative decision to change their behaviors."[90]

To be sure, choice architects normally design and implement their interventions in order to alter people *behaviors* 'in a predictable way', and the effectiveness of such interventions, as we saw in section 3.3, is therefore usually assessed by the extent to which they do. Indeed, this is just what my proposed distinction between internal and external routes to behavior change was meant to capture – i.e. to the extent that nudges are explicitly intended to *directly* affect people's actions and choices, they arguably *are* best regarded as external routes to compliance gaining.[91] Regardless of the nudger's original intentions, however, this fact leaves wide open the following possibility:

> Precisely *by* directly affecting the nudgee's *behaviors*, a given nudge may also, under certain favorable conditions, *indirectly* affect her *beliefs*.

This hypothesis, to my knowledge, has so far not received appropriate attention in the current nudging literature. This is at least surprising, as the long tradition of research in social psychology that I have been discussing throughout the present work seems to have by now uncontroversially established at least the following basic fact about human cognition: an exogenously-induced *behavior* change can easily bring about a *mind* change – i.e., a shift in the contents of people's self-attributed beliefs and attitudes. Behavioral scientists, on their part, often seem to just take this proposition for granted – to regard it simply as part of the received wisdom in their profession. To mention just one telling example, in discussing the advantages of exploiting the tools afforded by the emerging field of neuroeconomics to assess preferences directly – i.e., without relying on people behaviors or self-reports – Ariely and Norton (2008) begin their article by showcasing standard evidence that, quite routinely, both inside and outside the lab, people's actions do not

[89] Cf. Dennett (2013).
[90] Oliver (2017: 111, emphasis added).
[91] Cf. sections 1.4, and 3.4 above.

merely *reveal* or *reflect* their preferences, but rather *create* them. "Actions", in their own words, "are not merely the consequence but also the cause of preferences."[92] I hence find eminently plausible to suppose that, by first influencing actions, nudges might influence the formation of subsequent beliefs as well. When this is the case, I suggest, it would be appropriate to refer to the ensuing mental state as a *nudged belief*, which I propose to characterize as follows:

A *nudged belief* is a belief whose formation has been influenced by a nudge-induced action.

In order to clarify this notion, let me begin by taking note of a widely held view about the nature of beliefs. Most epistemologists today standardly construe such mental states as largely automatic responses to the available evidence, over the formation of which we typically do not have any direct, voluntary control.[93] What this means is simply that, once we have been exposed to the relevant evidence, the process of forming a corresponding belief is not to be thought along the lines of 'raising one's arm', but rather along the lines of 'digesting'. This is of course not meant to imply that we simply lack any control whatsoever over the formation of our beliefs. What the standard view suggests is rather that such control, when it is present, will only be of an indirect sort. To exemplify, upon entering a dark room, we cannot intentionally form the belief that the light is on. To be sure, however, we normally know which specific actions we need to perform in order to bring about the formation of such belief (e.g. opening the windows or reaching for the nearest light switch). In other words, while we cannot intentionally start to believe something – believe 'at will', as it is often put – we can certainly exercise various forms of indirect control over the formation of our beliefs by intervening on our environment. Needless to say, this same sort of intentional, indirect control over the formation of our beliefs may be exercised not just by ourselves, but by *others* as well. E.g., unless you are blind, if we enter the dark room together, I can obviously cause you to form the belief that the light is on by opening to the windows or reaching for the light switch myself. By the same token, if, while inquiring about a specific issue, you have failed to consider a crucial piece of evidence, and thereby come to form a false belief on the matter, I can certainly bring the relevant piece of evidence to your attention, and, by so doing, (hopefully) contribute to your forming a true belief instead. The very nature of our beliefs seems to clearly allow for such possibilities. Indeed, it seems only fair to maintain that most epistemologists today would find the above observations hardly in need to be seriously argued for. In discussing a different, though closely related topic, for instance, Grundmann (2021) observes the following:

[92] Ariely & Norton (2008: 15). It is also worth noting, for reasons that will become clear in what follows, that according to the authors preferences created by actions are not 'introspected', but rather 'inferred' from memories of past behavior.

[93] Cf., e.g., Alston (1989).

> "Obviously, we can intentionally trigger certain automatic belief-forming mechanisms in order to make people believe … certain propositions. We can thus *nudge* people into adopting certain doxastic attitudes."[94]

For our present purposes, then, the relevant point is that we not only have the ability to, but in fact quite routinely *do* influence other people's beliefs simply by manipulating their environment. Whenever we do this, we are in effect actively engaging in some form or other of what Neil Levy aptly calls *epistemic engineering*, and characterizes very broadly as the 'management of our epistemic environment'[95]. Indeed, a fundamental theme running through Levy (2022) is that, as it is well-known, Western epistemology has traditionally been heavily individualistic .[96] As a consequence, it has typically, yet to large extent misleadingly, construed knowledge acquisition as a primarily individual achievement. Unfortunately, however, this predominant focus on the knowing subject has led many to cultivate various manifestly unrealistic, and mostly scientifically unsupported, myths about the 'astonishing' powers of human individual, unaided cognition. Most importantly, it prevented many epistemologists from fully acknowledging the intrinsically social nature of such cognition, and to consequently underestimate the crucial role played by a vast array of social processes in the generation of human knowledge. Once we finally begin to think of knowledge as a deeply social phenomenon, according to Levy, epistemic engineering can be seen not just as a useful tool to improve belief formation, but also – at least amongst natural born influencers – as the standard route to belief revision. Here is how Levy summarizes the matter:

> "Change minds by changing the world: physical and social. That, I suggest, is not only how minds are most effectively changed. It is how minds have *always* changed."[97]

In the attempt to clarify the notion of a *nudged belief*, the above considerations have isolated a very general, and arguably quite common, epistemic phenomenon. The phenomenon in question, as we just saw, consists in some individual (or collective) agent other than ourselves intentionally manipulating our environment in order to influence our belief-formation processes. In honor of Levy's seminal contribution, let us therefore agree on referring to the very broad class of mental states generated through this external route as *epistemically engineered beliefs*. With this terminological stipulation in place, *nudged beliefs* can now be considered as a particular subclass of epistemically engineered beliefs. According to my proposal, that is, not all epistemically engineered beliefs are, ipso facto, nudged beliefs. Indeed, on the way in which I will henceforth use this expression, a nudged belief is not just any belief that has been formed as a consequence of someone else intervening on my environment for this purpose. In order for a nudged belief to count as such, said environmental manipulation has to eventuate in a specific *action* on my part – i.e. a *nudge-*

---

[94] Grundmann (2021: 3, emphasis added). It is worth noting, in this regard, that Grundmann coined the expression *doxastic nudging* in order to refer to various (mainly verbal) means to steer people's doxastic attitudes by 'triggering our biases in a smart way'. The typical cases that he has in mind are ones in which we can make people believe certain propositions by, e.g., 'framing them in especially persuasive ways' or 'presenting them as common ground'. Cf. *ibid.* p. 3. While finding such cases well worth of systematic epistemic investigation, the kind of *nudged beliefs* that I am here concerned with, as it should be clear from what follows, are not a result of 'doxastic nudging', at least not in the sense in which this expression is currently being used by its originator.

[95] Levy (2022, *Preface*, xv).

[96] Cf, e.g., O'Connor et *al.* (2024).

[97] *Ibid.* p. 84.

induced action. According to my proposal, it is my performing of this latter action that might subsequently lead to the formation of a nudged belief. The relation between the two classes of beliefs can hence be schematically represented as shown in Fig. 3.

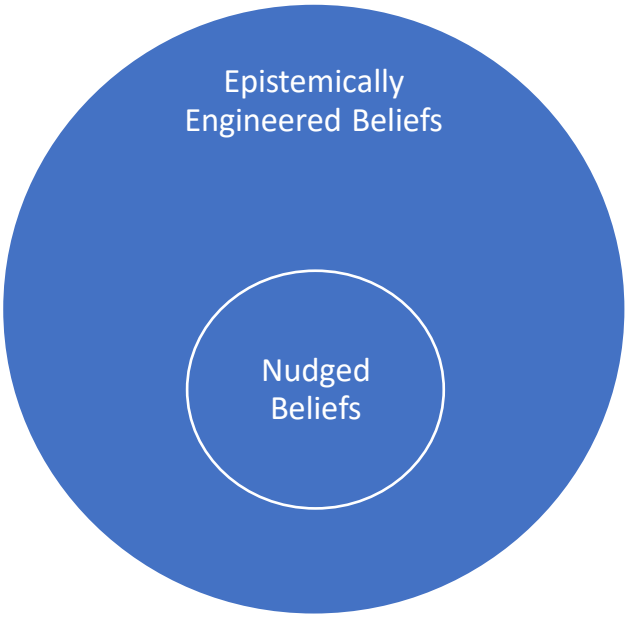

Fig. 3. Relation between *nudged beliefs* and *epistemically engineered beliefs*.

A nudged belief, then, is a particular kind of epistemically engineered belief, whose formation has been influenced by a nudge-induced action. What we still need, at this point, is a plausible, and empirically testable story about *how* this idea might work. The last step of my present case for nudged beliefs will accordingly consist in addressing the following question:

How might a *nudge-induced action* lead to the formation of a *nudged belief*?

An answer well worth exploring, I will suggest, is to be found in the current philosophical and psychological debate on *Self-Knowledge* – i.e. the knowledge of our own mental states (henceforth, SK) – and, in particular, in a specific family of accounts of this faculty often referred to as *Self-Interpretation* accounts. For present purposes, however, it will not be necessary – and nor it would be particularly useful – to provide a general overview of such debate.[98] Suffice it to say that the main objective of current research on SK is to provide a satisfactory explanation of humans' ability to come to know about their own mental states, such as, e.g., their beliefs, emotions, or preferences. The fundamental question that philosophers and scientists have been grappling with for ages is hence as follows: how – i.e. through which mechanisms – do we normally acquire such knowledge? In this regard, it is crucial for our purposes to note that, until relatively recent times, the standard way of thinking about this extremely complex issue rested on a seemingly rock-hard, and allegedly unquestionable theoretical assumption. This nearly indisputable proposition, according to scholars of all stripes, was that the cognitive process (or processes) by means of

[98] Two sufficiently detailed introductions to the current, and rather complex debate on Self-Knowledge are Gertler (2024), and Schwitzgebel (2024, section 2).

which we can normally come to know about any of our *own* knowable[99] mental states – however it might work – either is or must be profoundly different from the one that we instead normally rely on in order to come to know of the mental states of *others*. Indeed – or so the orthodoxy went – whereas of course we typically cannot but attribute mental states to *others* based on the observation and interpretation of their overt (physical and/or verbal) behavior and circumstances, it would simply be preposterous to even suggest that, barring a very limited set of highly unusual situations, a similar inferential route might indeed constitute the normal way in which we attribute mental states to *ourselves* as well. The largely predominant view, in particular, was that the kind of access we typically enjoy to our *own* mental states is of a *transparent* or *direct* sort – i.e., in Nisbett's words, a more or less 'direct readout' of our inner reality. To exemplify, the general idea would then be that the way in which I normally come to know that *Clara* loves sushi, and the one in which I normally come to know that *I* love sushi are, epistemically speaking, two entirely different beasts – whereas, for obvious reasons, I could only come to know that she likes sushi by paying attention to her overt behavior, and to the circumstances in which she acts (her choice architecture), it would simply be out of the question that I might ever come to know about my *own* food preferences by employing the same epistemic means, except perhaps in highly artificial lab settings or deeply unusual real-life situations.

To cut a (very long) story short, over the last several decades, this long-standing dogma has been systematically and radically questioned on both theoretical and empirical grounds, and this is where the above-mentioned Self-Interpretation (henceforth, simply SI) accounts of SK enter our scene. This family of accounts grew out of what some researchers – both in the philosophical and the psychological camp – began to perceive as a deep-seated dissatisfaction with the received picture of SK sketched above.[100] What the first supporters of such accounts were proposing, as a consequence, was to substitute the traditional, and then standard theoretical approach to SK, with an alternative picture of this characteristically human faculty. On the new picture they started advocating, both knowledge of our own mental states (SK) and knowledge of the mental states of others (Other knowledge), would in fact normally be acquired in a similar way. In particular, both SK and Other knowledge would commonly rely on basically the same type of cognitive mechanism (or mechanisms), and, crucially, this latter mechanism would not provide us with a *direct* or *transparent*, but rather with a largely *inferential* access to mental states in general. Another way to put this would be to say that, on the new picture, both SK and Other knowledge would, at the end of the day, rely on the very same kind of evidence. In the contemporary philosophical literature, this latter thought is usually traced back to the work of Gilbert Ryle, whose seminal treatment of the matter famously contains the following, much quoted, remark:

> "The sorts of things that I can find out about myself are the same … that I can find out about other people, and the *methods* of finding them out are much the same."[101]

In line with what Nisbett, as we saw, takes to be one of psychology's major insights into the way in which our minds work, the central claim of SI accounts is that acquiring SK – i.e. coming to

[99] This way of thinking, indeed, usually does not rule out the possibility that some of our mental states – e.g. such unconscious desires or fears – may well be in principle inaccessible to us.

[100] Cf., in particular, Ryle (1949), Bem (1967, 1972), Dennett (1991), Lawlor (2009), Carruthers (2009, 2011), Carruthers et *al.* (2012), and Cassam (2014).

[101] Ryle (1949: 155, emphasis added).

know about our own mental states – requires us to engage in a process of *self-interpretation*. This claim, in turn, takes us back to Daryl Bem's 'heretical' idea that we briefly considered in section 3.2 above while discussing the dissonance data. According to Bem, as you might recall, our mental lives would not be transparent to us. In particular, there would be plenty of situations in which introspection fails us, and we more or less automatically become the observers of ourselves in order to *infer* how we must think and feel about things – i.e., in effect, self-attribute mental states of various sorts – by relying on recollections of our own past behaviors and circumstances. When this is the case, according to Bem, the process (or processes) by means of which we attribute mental states to *ourselves* are the same ones that we normally rely on in order to attribute mental states to *others*. As we saw, he summarized the basic tenets of his Self-Perception Theory as follows:

> "Individuals come to "know" their own attitudes, emotions, and other internal states *partially* by *inferring* them from observations of their own overt behavior and/or the circumstances in which this behavior occurs. Thus, to the extent that internal cues are weak, ambiguous, or uninterpretable, the individual is functionally in the same position as an outside observer, an observer who must necessarily rely upon those same external cues to infer the individual's inner states"[102]

The reason why I requoted this passage is that it allows us to clarify a final important aspect of the hypothesis introduced earlier. A given nudge, according to my proposal, may *indirectly* affect our beliefs by *directly* affecting our behaviors. As I explicitly put it, however, this will be more likely to happen *under certain favorable conditions*. In light of the above, we can now be more specific about what such background conditions consist in. The conditions in question, in my view, are precisely the arguably very frequent ones in which, in Bem's words, 'internal cues are weak, ambiguous, or uninterpretable' – i.e. situations where our initial introspective search does not deliver any clear output. In illustrating the hypothesis, I used the example of someone who lacks any clear-cut attitude towards a given environmental issue, such as, e.g. the need to cut down on unnecessary air travel, or the benefits of recycling grey water for domestic uses. If queried about the issue, or if considering the question herself, this person would likely find herself in a situation in which the introspective search for an answer that she engages in fails to deliver any clearly interpretable signal. What Bem's well-supported theory predicts is that, on many such occasions, she will not just conclude that she simply cannot tell, nor admit to not having any opinion on the matter – she will rather start looking for cues in her recollection of her own past behaviors. It is especially on such occasions, I suggest, that a nudge-induced action on her part may well have an impact on the formation of her subsequent beliefs.

The scope of Bem's influential theory is very broad. Indeed, as the above quoted passage makes clear, the theory is intended to apply to 'internal states' in general. My own proposal, on the other hand, has a much narrower scope, as it only concerns what – I have argued – constitutes a specific subclass of our beliefs – i.e. nudged beliefs. In this regard, the most systematic, and empirically informed attempt at developing a SI account of SK to date is arguably due to Peter Carruthers, and it goes by the long name of *Interpretive Sensory-Access Theory* of SK – ISA Theory, for short.[103]

---

[102] Bem (1972: 2, emphases added).

[103] Cf. Carruthers (2011).

Now, as it was the case for the countless intricacies of the current debate on SK, a complete treatment of the many aspects and predictions of the ISA theory – let alone a global assessment of its merits and limits – would arguably take a dissertation of its own. Thankfully, however, this daunting task was never amongst our goals to begin with. The much humbler achievement that the present discussion was meant to accomplish, as you will recall, was that of coming up with a plausible, and empirically testable story about how nudged beliefs might work. In particular, to provide an experimentally treatable answer to the following question: how might a *nudge-induced action* eventually lead to the formation of a *nudged belief*? What we are after, then, is in effect, ideally, a psychologically plausible *model* of nudged beliefs. To this end, what matters for present purposes is that beliefs happen to feature prominently in Carruthers' ISA theory.[104] As a consequence, I believe that it provides valuable theoretical resources that one can profitably draw on in order to at least sketch such a model in its broad outlines. With this much more limited objective in mind, the ISA theory, as its author suggests, can usefully be seen as the conjunction of three basic theses.[105] According to the first one, our minds would only contain a single mental faculty – or *mindreading system* – charged with attributing beliefs both to ourselves, and to others. It follows that the normal – in fact, according to Carruthers, the only available – way to acquire knowledge of our own beliefs would be produced by, as it were, turning this single mindreading system on ourselves. Our coming to know our opinions about various matters, that is, would constitute a self-directed form of mindreading. According to the second thesis, the inputs to our mindreading system would all be *sensory* in characters – where the term 'sensory', however, has to be construed broadly enough so as to include all the forms of *perception* (such as, e.g., proprioception or interoception), as well as our visual and other forms of *imagery* (such as, in particular, inner speech). According to the last, and crucial thesis, we can only enjoy a non-interpretive or direct access to – i.e. we can only be immediately aware of – our own sensory mental states. As a consequence, Carruther's theory is committed to the claim that knowledge of all other forms of mentality, and hence, in particular, of our own *beliefs*, can only be *inferential*, and *sensorily-mediated* – whence its name: *Interpretive Sensory-Access Theory* of SK. On the ISA theory, then, our judgments about our own beliefs would be epistemically on a par with our judgments about others' beliefs, as both the *processes* by means of which we attribute beliefs to ourselves and others, and the kind of *evidence* that our minds rely on in order to perform this task would be essentially the same. In particular, in spite of the fact that both types of judgments often share a similar phenomenology, and feel like 'immediate readouts' of reality, on the present picture, our access to the non-sensory contents of our own minds would not be different, in principle, to our access to the non-sensory contents of other people's minds – i.e. it would rely on sensory information about past actions and circumstances, and it would therefore not be direct or transparent, but rather a matter of (usually unconscious) interpretation. For this reason, SI accounts such as Carruthers' are often

[104] Strictly speaking, the ISA theory does not directly target "standing attitudes", such as, e.g., *beliefs*, *memories*, or *intentions*, but rather our occurrent thoughts and thought processes, paradigmatic examples of which its author takes to be *judging*, *actively wanting*, and *deciding*. However, Carruthers himself takes beliefs to fall squarely within the scope of his theory, based on the hardly controversial assumption that "knowledge of our own standing attitudes *depends* upon knowledge of the corresponding (or otherwise suitably related) current mental events." Cf. *Ibid.* Preface, xi, emphasis added.

[105] Cf. *Ibid.*, Ch. 1. For the reasons mentioned in fn. 104 above, in presenting the ISA's basic tenets, I will couch them in terms of *beliefs*.

referred to in the current literature as *symmetric* or (Self/Other) *parity accounts* of SK.[106] It is finally important to note, in this regard, that this picture of SK does not at all rule out that, as it is indeed the case, we can normally come to know more about our own beliefs, than we can ever come to know about others' beliefs. The crucial point, however, in that, at least on the ISA theory, this plain fact has nothing to do with a difference in the way in which we acquire such knowledge. It is rather a straightforward consequence of the further, obvious fact that, lacking direct access to the sensory states of other people, we will always have *more* data available about ourselves, than about them. The fundamental idea, indeed, is that in attributing beliefs to ourselves and others we do not rely on the *same*, but rather on the same *kind* of evidence.

Regardless of whether or not one is willing to buy into Carruther's whole picture of SK, his ISA theory, as I put it above, arguably provides a host of valuable theoretical insights that one can profitably draw on as ingredients in order to sketch a model of nudged beliefs. So here, I propose, is a psychologically plausible, and empirically testable story about how, in very broad outline, nudged beliefs might work – i.e. about the general way in which a *nudge-induced action* might lead to the formation of a *nudged belief*.

Towards the beginning of the present section, as you may recall, I promised that, in due course, I would have revealed what, at least according to Richard Nisbett, constitutes psychology's second major insight into the workings of our minds. It is now time for me to honor that promise. The insight in question, in the words of the great psychologists, reads as follows: "the *situations* we find ourselves in affect our *thoughts* and determine our behavior far more than we realize"[107]. As it should be clear by now, some of those 'situations' happen to be ones in which our decisional environment has been intentionally designed and intervened on by choice architects in order to make it more likely that we will behave in a predictable way, while at the same preserving our freedom of choice. Nudge-based policies, that is, are typically put in place in order to gently induce us to freely and intentionally perform a vast array of quite specific *actions* – such as, e.g., voting for a candidate, selecting and eating a food, checking a box in a form, or tossing a disposable glass in the appropriate container. Based on what we said, then, each one of these specific actions, when it is actually performed by a nudgee, will count as *nudge-induced action* – i.e. while indeed free and intentionally performed, it is such as to satisfy the following counterfactual: if the relevant nudge had not been in place, the targeted individual would have been less likely to perform it. Now, to the extent that nudge-induced actions are typically *physical* actions, while we perform such an action, we will normally be *sensorily aware* both of performing it, and of the circumstances in which the action is being performed. Moreover, we will also be likely to form a belief about having performed the action in question, and to subsequently be able to retrieve this information from our memory. What all of this means, for our purposes, is that nudge-induced actions – just as any actions that we are (or can be) sensorily aware of performing, and about which we can form any memories at all – are bound to engage our mindreading system. Now, on the ISA theory of SK – and, partly, on other SI accounts of this faculty, such as, e.g. Bem's Self-Perception Theory – our ability to self-attribute beliefs would be ultimately grounded in a perceptual awareness of our own circumstances and behavior. It hence seems reasonable to expect that, after having performed a nudge-induced action, an individual will be more likely to

[106] Cf., e.g., Schwitzgebel (2024, sec. 2). Parity accounts have been defended by, e.g., Nisbett & Ross (1980), and Gopnik (1993).

[107] Nisbett (2015: 15, emphases added).

self-attribute beliefs that she perceives as congruent with her own past action – i.e. to form a nudged belief.

Before drawing some conclusions, however, let me first clarify an important aspect of this proposal. Differently from the ISA theory, the above story, as it stands, is not committed to the assumption that we can *only* have inferential access to our beliefs, nor to the assumption that we can have inferential access to *all* of our beliefs. In other words, following most accounts of SK, the model leaves quite open the possibilities that there may in fact exist more than one kind of access to some of our beliefs, as well as that there may be some of our beliefs access to which is for us not normally inferential.[108] Indeed, this simply follows from my earlier claim about the *favorable* background *conditions* under which a given nudge, in my view, may *indirectly* affect our beliefs by *directly* affecting our behaviors. Said conditions, I suggested, will be the arguably very frequent ones in which our internal cues – i.e. the deliverances of our introspective search – are weak, ambiguous, or uninterpretable. The model, in other words, clearly makes room for the existence of cases in which the outputs of our introspection will instead be strong, clear, and perfectly interpretable. It only suggests that this latter kind of cases are not the most favorable ones with respect to the formation of nudged beliefs.

Needless to say, the one just sketched is only a skeletal model of nudged beliefs. In particular, it is admittedly very far from constituting a complete mechanistic explanation of its intended target system. It is hence much more plausibly regarded as what Craver (2006) refers to as a *mechanism schema* – i.e. a description that abstracts away 'to a greater or lesser extent' from the specific details of any particular mechanism.[109] Yet, or so I like to think, it is a model nonetheless, and models can play many roles in science other than fully explaining things. In particular, they can often be relied on as a kind of useful *heuristics* for designing future experiments. This latter function, as I will shortly suggest, is indeed the one that my proposal was intended to fulfil. In this regard, there is an obvious reason why I so far intentionally avoided calling my model an 'explanation', and more causally referred to it simply as a 'story'. The reason why this latter term seems definitely more appropriate is that, in order to more or less completely *explain* a given phenomenon, one must of course already possess solid grounds for believing that the phenomenon in question *exists* in the first place. And I don't! Indeed, the final question that needs to be addressed, at this point, is arguably the following: is there any empirical evidence for the existence of the mental states that I called *nudged beliefs*? Or, to put it even more bluntly: is nudge-induced belief change really a thing? At present, unfortunately, I simply do not have an answer to this question. What I *do* have, however, is a very simple model – indeed, the kind of theoretical entity that Craver (2006) would perhaps somewhat dismissively refer to as a *how-possibly* model[110] – that can arguably be used as a guide for designing a psychological study intended to find out. As a matter of fact, I *did* design and perform a pilot study which, to my eyes at least, constitutes a valid first stab at experimentally investigating the matter, and I will report on its general structure and main results in the next and final chapter of this dissertation.

---

[108] Schwitzgebel (2024), for instance, does not shun resorting to the 'surely operator' to express this thought: "Surely there is more than one process by means of which we can obtain self-knowledge". Cf. Schwitzgebel (2024, sec. 2).

[109] Craver (2006: 360).

[110] Ibid., pp. 361-62.

4. The Uniurb Study

4.1 Introduction

Although public policies inspired by psychological intuitions have likely been around for as long as mankind existed, the turn of the last Century introduced us natural born influencers to a new policy game in town: *nudging*. The good news coming from research on human judgment and decision-making is indeed that *directly* intervening on what people think and feel by means of persuasion is not always necessary to affect the way in which they act. In particular, as we have seen throughout the present dissertation, evidence shows that it is possible to intentionally alter people's behavior – i.e. make it more likely that they will act in a certain way – without resorting to coercion, but rather by simply reshaping their decisional environment or 'choice architecture'. According to its original characterization, indeed, a *nudge* is any aspect of our choice architecture that can influence our behavior in a predictable way without changing our options or economic incentives, while at the same time preserving our freedom of choice.[1] The rationale of nudge-based policies, then, is that the quality of our decisions can be demonstrably improved simply by changing the way in which the same options are presented to us. In spite of the many ethical issues raised by their widespread use[2], many policymakers at present regard nudges as a very profitable instrument to help people make choices that they themselves, on reflection, would consider better. As we saw[3], nudges have so far been successfully applied to many different policy domains – such as e.g. politics, finance, and health. As we have also had occasion to discuss[4], they have proved particularly efficacious in the environmental domain, by encouraging the adoption of various *climate-change-responsive behaviors* (CRBs) – such as, e.g., conserving energy, recycling, keeping air travel at a minimum, or driving fuel efficient cars. For this reason, most governments and institutions today regard so-called *green nudges* – i.e. pro-social nudges whose main goal is to reduce negative environmental externalities – as a key tool in combating climate change, both on the mitigation and adaptation front. In particular, we said, green nudges appear very promising to bridge the so-called *Value-Action Gap* – i.e. the large disconnect between our possession of environmental knowledge and awareness, on the one hand, and our adoption of CRBs, on the other – by helping people to realign their environmental decisions with what they already appear to regard as the right thing to do.

The psychology of nudging

Nudges are the relatively recent offspring of a long called-for marriage between economics and psychology.[5] A natural consequence of this fact is that serious research on nudge-based policies, while currently thriving, is admittedly still in its infancy. In particular, its highly interdisciplinary

[1] Thaler & Sunstein (2021: 8).
[2] Cf. Ch. 2 above.
[3] Cf. section 1.4 above.
[4] Cf. 1.5 and Ch. 2 above.
[5] Cf. sections 1.4 and 3.3 above.

nature makes it the case that its subject matter can be approached from many different, and mutually reinforcing, disciplinary angles. In this regard, many today feel that current research on this topic would stand to greatly profit from engaging more systematically with the *psychology of nudging* – i.e. pursuing research explicitly aimed at exploring, both theoretically and empirically, nudges' cognitive underpinnings. On a theoretical level – as we saw for the case of defaults[6] – the same observed behavioral effect will indeed typically allow for different explanations, each pointing to a different psychological mechanism. This theoretical state of affairs has far-reaching practical consequences, as it severely affects nudges effectiveness and replicability. Indeed – as we illustrated with the case of indirect water conservation[7] – nudges' effectiveness is extremely context-dependent, and a given nudge which proved remarkably effective in one context, can prove largely ineffective (or worst, backfire) in others. The natural cure for this disease, some have forcefully argued, consists in opening nudges' black box, and start looking for mechanisms – i.e. psychological entities and processes that undergird nudges' effectiveness, and can hence be credited for actually making a difference to the behavioral variable of interest. The main goal of the present work is to contribute to this ongoing search for mechanisms. Its guiding idea is that the long-standing tradition of research on attitude-behavior links in social psychology, and the fledgling research on nudge-based policies can helpfully illuminate each other. In particular, as I argued in the last chapter, I believe that a mature psychology of nudging could profit immensely from building on the already existing, hardly-acquired, and empirically well-supported knowledge of the many ways in which our attitudes and behaviors are causally linked in reciprocal ways. One indication that this is indeed the case, as we saw, consists in the fact that, contrary to what one might be tempted to casually assume, beliefs *matter* to nudging – i.e. which particular beliefs the nudgee happens to hold (e.g. her ideological priors) has been shown to make a sizable difference to the effectiveness of the nudging process.[8]

A self-interpretation model of nudged beliefs

A further issue in the psychology of nudging that seems well worth exploring is the extent to which the nudging process might have a measurable impact on people's self-attributed beliefs. Since this possible consequence of nudging has so far received surprisingly little attention in the relevant literature, the specific goal of this dissertation has consisted in attempting to redress this imbalance. Indeed, as I argued, to the extent that, in general, an exogenously-induced *behavior* change can easily bring about a *mind* change – i.e., a shift in the contents of people's self-attributed mental states – it appears perfectly reasonable to expect that the nudging process may have an impact on people's beliefs, over and above their behaviors.[9] As a consequence, in section 3.5 I considered the hypothesis that precisely by *directly* affecting the nudgee's behaviors, a given nudge may also, under certain favorable conditions, *indirectly* affect her beliefs. In particular, I introduced the term *nudged beliefs* in order to refer to beliefs whose formation has been influenced by a *nudge-induced action*, and I put forward a specific proposal concerning how this process might unfold – i.e. through which mechanisms a nudge-induced action might lead to the formation of a nudged

[6] Cf., in particular, section 3.3 above.
[7] Cf. section 3.3 above.
[8] Cf. section 3.4. above.
[9] Cf. section 3.5 above.

belief. The crucial mechanism, according to my model, is one of Self-Knowledge – i.e. knowledge of our own mental states. In this regard, the central claim of so-called Self-Interpretation accounts of this faculty, as we saw, is that coming to know what we think or feel about various matters often requires us to engage in a process of self-interpretation.[10] What this means is that, on many occasions, we will more or less consciously self-attribute mental states based on an observation of our own past behavior, and the particular situation in which that behavior occurred. My model, or mechanism schema, builds on this insight and proposes that the formation of a nudged belief might come about in the following three steps. (1) A *nudge* will typically induce us to freely and intentionally perform a specific *action* (e.g. tossing a disposable glass in the appropriate recycling bin); (2) performing this action will lead to the formation of a memory (and corresponding belief) about having performed the action in question, as well as about the particular situation in which it was performed; (3) this memory will trigger a process of self-interpretation the output of which will be the self-attribution of a belief that we perceive as congruent with our own immediately past action – i.e. a nudged belief. This last step, according to the model, will be particularly likely to take place when our internal cues – i.e. the deliverances of our introspective search – are weak, ambiguous, or uninterpretable. This mechanism schema can hence be represented as shown in Fig. 4, where nudge-induced actions are referred to simply as *nudged actions*.

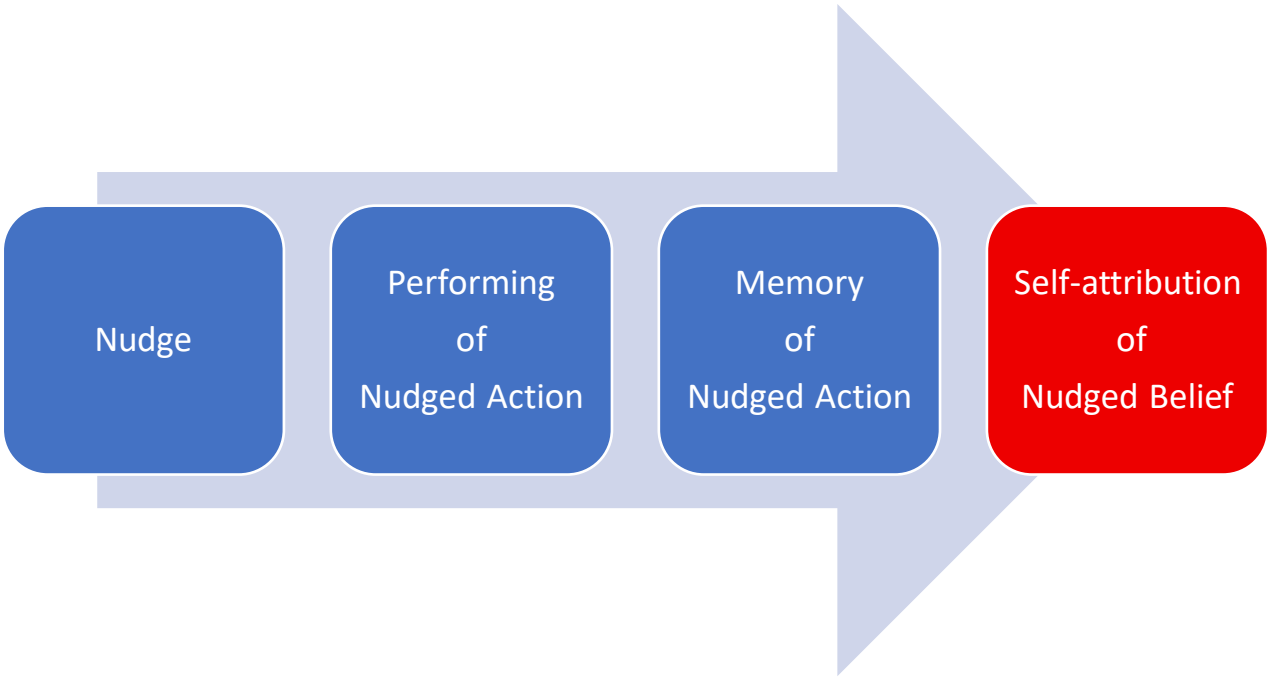

Fig. 4. Mechanism schema of the formation of nudged beliefs.

In the present chapter, I will present and discuss the results of a relatively simple pilot study that I designed and performed in order to begin testing the above self-interpretation model of nudged beliefs, and which I hope might constitute a valid starting point in order to develop future, more

[10] The two self-interpretation accounts that inspired my proposal, as we saw in section 3.5, are Daryl Bem's (1972) Self-Perception theory, and Peter Carruthers' (2011) ISA theory, respectively.

elaborate experimental studies aimed at investigating the effects of nudge-based policies on the formation of people's beliefs. As the present dissertation focuses on *environmental* beliefs and behaviors, the initial inspiration for the following experiment came from an influential seminal study on the effects of behavioral information on reported environmental attitudes due to two prominent social psychologists: Shelly Chaiken and Mark Baldwin.[11] Chaiken and Baldwin's (1981) study was explicitly designed in order to test specific consequences of Daryl Bem's (1972) Self-Perception theory – discussed in chapter 3. Their main finding is that subjects with poorly defined prior environmental attitudes will display a marked tendency to express either pro- or anti-ecology attitudes as a function of which kinds of past environmental behaviors – pro-ecology or anti-ecology – they were experimentally induced to recall.[12] In Chaiken and Baldwin's sophisticated study, the crucial independent variable – i.e. subjects' memories of their past behaviors – was manipulated by means of a questionnaire explicitly designed in order to make salient to the subjects either their past pro-ecology or their past anti-ecology behaviors. Since the present work is meant as a contribution to the psychology of nudging, the basic idea that guided the design of the following study was that subjects' past environmental behaviors could have been made salient to them, not by means a questionnaire, but rather by means of a green nudge.

The hypothesis

The self-interpretation model of nudged beliefs postulates that – especially when internal cues are weak, ambiguous, or uninterpretable – memories of past nudged actions can trigger a self-interpretation process the output of which will be the self-attribution of a belief whose content the individual perceives as congruent with her own past action. As a consequence, I predicted that exposing individuals to a green nudge will have a positive influence on their self-reported environmental beliefs. In particular, I expected that individuals who have been nudged to perform a pro-environmental action will exhibit more pro-environmental (post-manipulation) beliefs on a currently standard scale. Conversely, individuals who have not been nudged to perform the pro-environmental action in question will exhibit less pro-environmental (post-manipulation) beliefs as measured on the same scale.

The present study

To begin testing my hypothesis concerning the impact of green nudges on the formation of environmental beliefs, I conducted a pilot study at the University of Urbino. The study targeted the recycling behavior and environmental beliefs of a sample composed of first- and second-year undergraduate students enrolled at that university. The self-reported environmental beliefs of subjects in the treatment group were measured using the NEP Scale[13] both three weeks before, and immediately after exposing them to a recycling nudge. The nudge in question consisted in a perceptually salient variation in the number, dimensions, and specific location of recycling bins,

[11] Cf. Chaiken & Baldwin (1981).

[12] Chaiken and Baldwin assume that a measure of the structural consistency between the affective and cognitive component of a subject's environmental attitude will appropriately index the degree to which that individual possesses a well-defined (as opposed to poorly-defined) environmental attitude – i.e. the higher an attitude's affective-cognitive consistency, the better defined it will be.

[13] Dunlap et *al.* (2000).

coupled with a visual message appealing to a descriptive social norm concerning recycling at the University of Urbino. Exposing subjects to this recycling nudge was intended to induce them to perform a specific target action – i.e. toss a disposable drinking cup in the appropriate recycling bin. The recycling behavior of all subjects was observed and noted in order to assess whether subjects in the treatment group who actually performed the nudge-induced action would exhibit more pro-environmental (post-manipulation) beliefs as compared to the subjects in the control group.

4.2 Methods

Participants and design

The subjects in the study were either first- or second-year undergraduate students enrolled at the University of Urbino, and majoring in twelve different fields.[14] The students were recruited on a voluntary basis by means of twelve short presentations held by the experimenter during their classes between the months of December 2023, and January 2024. The presentations were meant to illustrate the general purpose, and outline the procedure of the study. However, in order to reduce reactance, the details of the procedure were not disclosed in full. In particular, during the presentations (and also – for the students who actually participated in the study – upon signing the informed consent form before taking part in the study's first session), the students were told that the study would require them to complete a questionnaire intended to acquire information about their environmental opinions, which they would have been asked to complete twice on two different occasions for measurement stability reasons. The students, however, were not told that, before participating in the second session, some of them (the treatment group) would have been exposed to a recycling nudge. Upon hearing the presentations, students were given the chance to add their names to a list to indicate their potential availability to participate in the study. Following comparable studies[15], the recruitment phase aimed at a sample size of at least 100 subjects. Unfortunately, while approximately 180 students – out of the roughly estimated 800 who heard the presentations – initially added their names to the lists, after repeated reminder emails, only 48 of them participated in the study (females: 70.83%, males: 27.08%, non-binary: 2.08%. Age: M = 21.2, SD = 4.2). All 48 subjects satisfied the other eligibility criteria for being included in the study – i.e. being an Italian native speaker, and not having been diagnosed with neurological or neurodevelopmental disorders. The study employed a 2 x 2 mixed-design ANOVA framework. The within-subject factor was *time* (pre-treatment vs. post-treatment), and the between-subject factor was *group condition* (treatment vs. control). The design allowed for the evaluation of main effects for time and group condition, as well as their interaction, to assess the

[14] The twelve majoring fields (Italian: *corsi di laurea*) were the following: "Scienze della formazione primaria" (primary education studies), "Scienze politiche, economiche e del governo" (political science), "Geologia per la sostenibilità ambientale" (environmental geology), "Economia e management" (economics and management), "Scienza della nutrizione" (nutritional science), "Informazione, media, pubblicità" (media studies), "Biotecnologie" (biotechnologies), "Sociologia e servizio sociale" (sociology), "Scienze umanistiche" (humanities), "Scienze biologiche" (biological sciences), "Lingue e culture moderne" (modern languages and cultures), "Scienze motorie, sportive e della salute" (motor, sport and health sciences).

[15] Cf., e.g., Chaiken & Baldwin (1981), Albarracín & Wyer (2000), Zani & Prati (2012).

treatment's impact on the dependent variables. The dependent variables were: *New Ecological Paradigm Scale* (NEP), waste sorting item (WS), and packaging item (P).

Measures and materials

Subjects' environmental attitudes were measured by means of a self-report questionnaire, answers to which were provided by participants on a five-point Likert scale. The questionnaire consisted in 17 items. The last two items have been explicitly designed to assess subjects' attitudes toward recycling. The first 15 items constitute the *New Ecological Paradigm Scale* (NEP Scale).[16] The NEP Scale was developed by sociologist Riley Dunlap and his colleagues to assess the degree of endorsement of an ecological worldview, and demonstrated good internal validity ($\alpha = .83$). It is currently the most widely used measure of environmental beliefs in studies that – such as the present one – use theoretical models for predicting environmental beliefs and behavior.[17] The NEP Scale has further been found to be useful in predicting both observed and reported behavior.[18] The 15 items of the NEP scale had already been translated into Italian by Gabriele Prati and Bruna Zani,[19] who generously allowed to make use of their translation for the present study. The questionnaire was administered using Google Forms, and can be found in the Appendix to the present chapter (section 4.5 below).
Following other studies[20], the nudge used to influence the behavior of subjects in the treatment group consisted in a perceptually salient variation – between treatment and control – in the number, dimensions, and specific location of the recycling bins, coupled with a visual message appealing to a descriptive social norm concerning recycling at the University of Urbino. A study by Andrews and her colleagues (2013) suggests that two- or three-compartment dedicated bins located next to the undifferentiated waste bin may be the best option to increase recycling accuracy. Two studies conducted by O'Connor and his colleagues (2010), and Miller and his colleagues (2016) respectively further suggest that placing recycling bins close to the point of consumption is a promising procedure to achieve the same goal. As a consequence, in the treatment condition a recycling collection unit including an undifferentiated waste bin, and composed of four joint bins in total (respectively labeled "plastic", "paper", "glass", and "undifferentiated waste") was placed in close proximity of the point of consumption. The bins composing the unit were much bigger than the ones used in the control condition, and placed next to a workstation sitting at which students were asked to complete their questionnaires, which supposedly made them rather easy to locate if looked for. A sign that read "Uniurb ricicla!" ("The

[16] Dunlap et *al.* (2000). The *New Ecological Paradigm Scale* constitutes a revision of the previous *New Environmental Paradigm Scale*. Cf. Dunlop & Van Liere (1978). Cf. Dunlap (2008) for an account of how the NEP Scale was originally conceived, and subsequently revised, Stern et *al.* (1995), for its connections with social-psychological theories, and Hawcroft & Milfont (2008) for a meta-analysis of 69 studies using the NEP scale in 36 different countries. The 15 items of the revised NEP Scale can be found in the Appendix to the present chapter (section 4.5 below).
[17] Cf., e.g., Stern et *al.* (1995), Stern et *al.* (1998), and Schultz & Zelezny (1998).
[18] Cf., e.g., Casey & Scott (2006), and Olli et *al.* (2001).
[19] Prati & Zani (2012).
[20] Cf., e.g., Andrews et *al.* (2013), O'Connor et *al.* (2010), Miller et *al.* (2016), Cosic et *al.* (2018).

University of Urbino recycles!") was placed on the wall, over the recycling unit.[21] In the control condition, the recycling bins were only three (respectively labeled "plastic", "paper", and "undifferentiated waste") and they were placed farther away from the point of consumption. They were also remarkably smaller than the ones used in the treatment condition, and placed in two different locations far removed from the workstation, which supposedly made them harder to immediately locate, if looked for. The recycling unit and the bins used for both experimental conditions are shown in Fig. 5a (treatment), and Fig. 5b (control) below.

Fig. 5a. Recycling collection unit used in the treatment condition.

[21] For the behavioral impact of descriptive norms in the context of littering and (curbside) recycling, cf. Cialdini et *al.* (1990), and Schultz (1999), respectively. Cf. Farrow et *al.* (2017) for a review of the effectiveness of social norms in encouraging pro-environmental behavior.

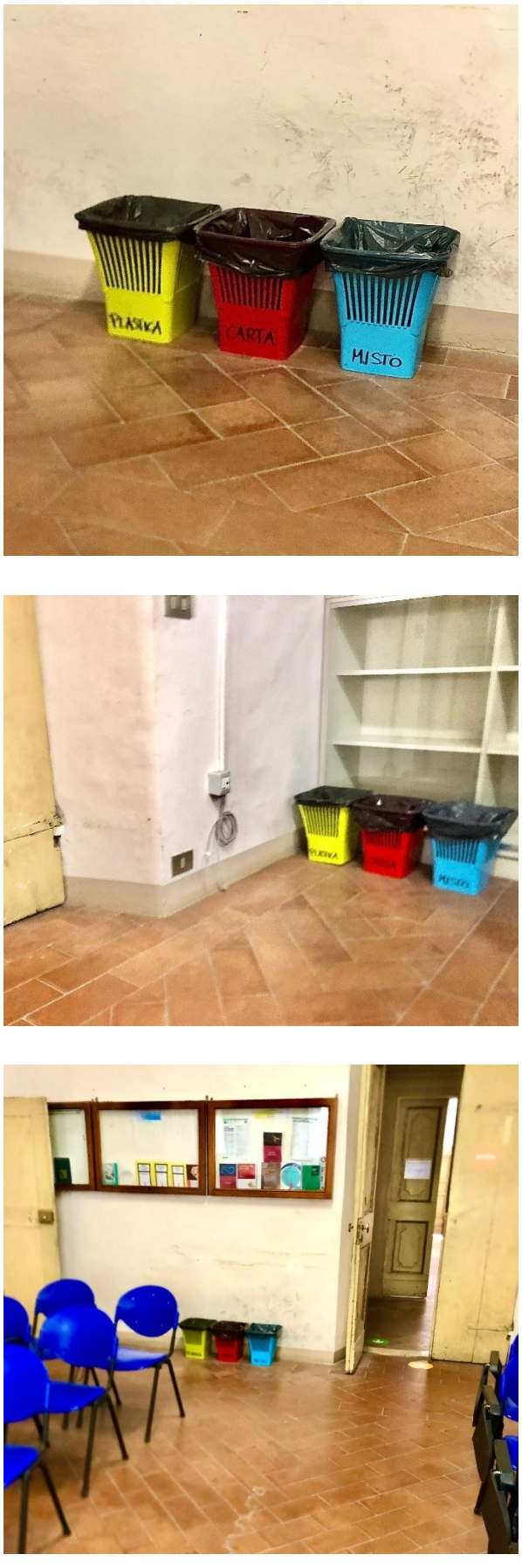

Fig. 5b. Recycling bins used in the control condition.

Procedure

The study was conducted over five weeks between April and May 2024 in a lecture room located within a building of the University of Urbino. It consisted in two experimental sessions (Session 1 and Session 2) which took place within the first and last week, respectively. Session 1 lasted approximately 5 – 10 minutes per participant. Session 2 lasted approximately 10 – 15 minutes per participant. Participants were randomly assigned to either the treatment or control condition based on the order in which they had accepted to participate in the study. Participants in both conditions completed the questionnaire twice – i.e. once per experimental session. In session 2, participants in the treatment condition were exposed to the recycling nudge. The procedure followed during the two experimental sessions was as follows:

Session 1. Upon entering the room, participants in both conditions were welcomed by the experimenter, and kindly asked to fill out and sign the informed consent form, which they all did. Participants in both conditions were then assigned an identification number, and they were kindly asked to begin completing the self-report questionnaire by entering their identification number on the first page of their questionnaire. Once they completed the questionnaire, participants were thanked, told that they would soon receive an email from the experimenter informing them about the time and date of the second session, and dismissed.

Session 2. Upon entering the room, participants in both conditions were welcomed and warmly thanked by the experimenter for having kindly accepted to take part in his study. They were then handed a disposable drinking cup, and cheerfully invited to help themselves to one of the soft drinks that had been placed on a table next to the workstation in order – they were told – to "celebrate together the end of the study". All participants accepted the offer, and poured themselves a glass of soft drink. Participants were then reminded of their identification number, and kindly asked to complete the self-report questionnaire for the second time. Upon having consumed the drink, participants had the chance of exhibiting their recycling behavior by either performing or not performing the target action – i.e. toss the disposable drinking cup in the appropriate recycling bin. The recycling behavior of all participants was observed and noted by the experimenter.

Data analysis

The primary goal of the analysis was to evaluate the effect of the recycling nudge (treatment) on participants' questionnaire scores across the following dependent variables: NEP-pre/NEP-post, WS-pre/WS-post, and P-pre/P-post. Before conducting the main analyses, these variables were Winsorized using the Winsorize function from the DescTools package (default options), in order to reduce the influence of extreme outliers. Mixed-design ANOVAs (repeated measures ANOVAs with a between-subject factor) were then performed for each dependent variable, with *time* (before and after treatment) as the within-subject factor, and with *group condition* (treatment vs. control) as the between-subject factor. These analyses allowed to examine the main effects of both time and condition, as well as their interaction, in order to determine whether the recycling

nudge (treatment) had a differential impact over time for the treatment and control groups. The significance level was set at $p < .05$ for all statistical tests. Statistical analyses were conducted using R.[22]

4.3 Results

Descriptive statistics for the key variables (NEP, WS, and P) are presented in Table 1 below, alongside a correlation matrix (Table 2 below) showing the relationships between these variables. For both tables, the scores are based on the mean of each participant across the two measurements – i.e. before and after being exposed to the recycling nudge (treatment).

To assess the efficacy of the recycling nudge, we conducted a chi-square test of independence to determine whether the nudge had a significant effect on participants' recycling behavior. The observed and expected frequencies of participants who recycled or did not recycle, separated by treatment group (treatment vs. control), are displayed in Table 3.

The chi-square test revealed a significant association between exposure to the nudge and recycling behavior, $\chi^2(1, N = 48) = 5.87, p = .015$. Specifically, participants in the treatment group were more likely to recycle their disposable drinking cups (86.4%) compared to those in the control group (53.8%). This result supports the conclusion that the nudge effectively increased recycling behavior among participants.

To examine the effects of time (before and after) and group condition (treatment vs. control) on participants' scores, separate mixed ANOVAs were conducted for each dependent variable: NEP-pre/NEP-post, WS-pre/WS-post, and P-pre/P-post.

For NEP-pre/NEP-post, the mixed ANOVA revealed a significant main effect of *condition*, $F(1, 90) = 10.13, p = .002$, indicating that participants in the treatment and control groups had significantly different overall scores, with higher NEP scores for the control group. However, the main effect of *time* was not significant, $F(1, 90) = 0.02, p = .88$, suggesting no significant change from before to after the treatment. Additionally, the interaction between condition and time was nonsignificant, $F(1, 90) = 0.02, p = .90$, indicating that the change in NEP scores over time did not differ between the treatment and control groups.

For WS-pre/WS-post, the mixed ANOVA showed no significant main effect of *condition*, $F(1, 90) = 0.07, p = .80$, and no significant main effect of *time*, $F(1, 90) = 1.26, p = .27$. The interaction between condition and time was also non-significant, $F(1, 90) = 0.32, p = .57$. These results suggest that neither time nor group condition had a significant impact on WS scores.

Similarly, for P-pre/P-post, there were no significant main effects of *condition*, $F(1, 90) = 0.05, p = .82$, or *time*, $F(1, 90) = 0.26, p = .61$. The interaction between condition and time was also not

[22] R Core Team (2025).

significant, $F(1, 90) = 0.37$, $p = .55$, indicating no significant differences between groups or changes over time for P scores.

Table 1. Descriptive Statistics

Variable	Mean	Median	SD	Range
NEP	3.45	3.80	0.67	2.00 - 4.80
WS	3.78	4.50	0.72	2.10 - 5.10
P	4.12	3.50	0.85	2.30 - 5.90

Note. $N = 48$. NEP = New Ecological Paradigm; WS = waste recycling item; P = Packaging item. The statistics shown in this table have been computed using the average, for each participant, of the pre-treatment and post-treatment measure.

Table 2. Correlation matrix

Variable	1	2	3
1. NEP	1.00	0.11	-0.02
2. WS	0.11	1.00	0.49**
3. P	-0.02	0.49**	1.00

Note. $N = 48$. Correlations marked with ** indicate significance at $p < .01$. The statistics shown in this table have been computed using the average, for each participant, of the pre-treatment and post-treatment measure.

Table 3. Recycling behavior by treatment condition

Recycling Behavior	Treatment Group: Yes	Treatment Group: No	Row Totals
Recycle (Yes)	19 (15.12) [0.99]	14 (17.88) [0.84]	33
Recycle (No)	3 (6.88) [2.18]	12 (8.12) [1.85]	15
Column Totals	22	26	48

Note. Expected frequencies are reported in parentheses; chi-square contributions are reported in brackets.

4.4 Discussion

The main goal of the present pilot study was that of taking a first step toward testing the self-interpretation model of nudged beliefs put forward in sections 3.5, and 4.1 above. According to this theoretical framework, memories of nudge-induced actions can trigger a self-interpretation process the output of which will be the self-attribution of a belief whose content we perceive as congruent with the action in question. The model hence predicts that exposure to a green nudge will have a positive impact on an individual's self-reported environmental beliefs. As the present study targeted the recycling behavior and the environmental beliefs of students enrolled at the University of Urbino, the hypothesis that it was explicitly designed to test was that subjects in the treatment group – having been exposed to a recycling nudge – would have exhibited more pro-environmental (post-manipulation) beliefs than subjects in the control group. The study did not find evidence for this effect. A significant difference was observed between the treatment and control groups with respect to NEP scores, with the control group scoring higher overall. However, no significant changes were observed over time for NEP scores, nor were there any effects for WS (waste sorting item) or for P (packaging item) scores across time or between conditions. These results hence suggest that, while group condition influenced NEP scores, the manipulation did not produce measurable changes in self-reported environmental behaviors, nor in environmental beliefs, raising questions about the treatment's efficacy.

A first aspect of the above results worth noting is that – as shown in Table 1 above – NEP scores were rather high in the target sample (Mean = 3.45). This is largely consistent with various findings coming from the body of research on environmental attitudes reviewed in Gifford and Sussman's (2012). In particular, two key demographic variables might have been responsible for the aspect in question – i.e. age, and gender. The subjects in the sample were indeed almost all young (Age: M = 21.2, SD = 4.2.), and predominantly women (Females: 70.83%), two categories that usually show higher levels of environmental concern.[23] It is worth noting, in this regard, that, in spite of their higher levels of environmental concern, women have also often been shown to exhibit lower levels of pro-environmental behavior and environmental knowledge than men.[24] Whereas some authors have suggested that the observed lower levels of environmental knowledge in women may be related to a lack of social encouragement to pursue scientific studies, some others have connected women's higher levels of environmental concern to higher levels of altruism and concern for health and safety.[25] It therefore seems plausible to conclude that high NEP scores in the target sample may be due to its composition.

A second, and arguably more relevant, aspect of the above results is that – as shown in Table 2 above – the NEP scores have failed to prove a good predictor of subjects' self-reported recycling behaviors. Scores on the NEP Scale are indeed either weakly or not at all correlated to scores on the items corresponding to the other two dependent variables – i.e. WS (waste sorting item) $r =$

[23] Cf, e.g., Honnold (1984) for young people, and Blocker & Eckberg (1997) for women.

[24] Cf., e.g., Arcury et *al.* (1987), and Stern et *al.* (1993). While several studies support this claim, it seems only fair to point out that, by Gifford and Sussman's own admission, as of 2012 research on gender and environmental concern was somewhat outdated, and in need to be revisited.

[25] Cf, e.g. Davidson & Freudenburg (1996), and Dietz et *al.* (2002).

0.11, and P (packaging item) r = - 0.02. On the one hand, this finding seems broadly consistent with the idea of a pervasive *Value-Action Gap* in the environmental domain discussed in section 1.5 above. On the other, it is however hardly consistent with what seems to be the prevailing attitude toward NEP's predictive power amongst researchers. Let us hence briefly consider both points in turn. Starting from the former, as we saw in section 3.1, the so-called *principle of cognitive consistency* – according to which individual's decisions and actions will normally be consistent with their opinions – played a central role in most psychological theorizing about attitudes' influence on behaviors from the 1950s on. As of today, the general assumption seems to be that attitudes are indeed a fairly reliable guide to behaviors in many domains, and the experimental investigations of social psychologists are rather aimed at identifying a vast array of moderating variables that can be shown to affect consistency by either increasing or decreasing the extent to which attitudes can reliably predict behavior.[26] As discussed in section 1.5, however, *environmental* attitudes seem to constitute an exception to the above-mentioned general assumption, as they have often been observed to be only weakly predictive of pro-environmental behaviors. Perhaps due to a well-documented social desirability bias, many people indeed exhibit higher levels of environmental concern than are actually reflected in their observed or self-reported behaviors.[27] As to the second point – i.e. the finding's inconsistency with the prevailing attitude toward NEP's predictive power amongst researchers – two aspects deserve mention. The first one is that, in general, attitudes' predictive validity has been repeatedly found to be affected by measurement issues. Indeed, as I put it in section 3.1, in order for people's attitudes to reliably predict their behaviors, there usually has to be a high level of correspondence, in terms of specificity, between the measures of the two variables – an observation usually referred to as the *principle of compatibility*. As I pointed out on that occasion, however, this does not mean that general attitude measures – such as the NEP Scale – are *never* useful in order to predict people's behaviors. On the contrary, some evidence suggests that global attitude measures can at times be useful in predicting recurring behaviors, such as, indeed, recycling.[28] In this regard, the second aspect worth mentioning is that numerous studies have found significant relationships between the NEP Scale and various types of behavioral intentions, as well as between the NEP Scale and both self-reported and observed behaviors.[29] Some of these studies, in particular, have specifically targeted recycling behavior. A study by Vining and Ebreo (1992), for instance, found that, although specific – as opposed to global – attitudes towards recycling accounted for a greater percentage of variance in subjects' recycling behavior, the global pro-environmental attitudes of subjects who did recycle were nonetheless stronger than the ones of subjects who did not.[30] A further study by Schultz and Oskamp (1996) found a strong positive relationship between subjects' level of environmental concern, and the amount of effort they were willing to exert to recycle.[31] In light of these studies, the above finding remains more difficult to account for. One possibility, as in the previous case,

[26] Some of those factors, as we saw, are the strength with which a given attitude is held, its value-expressive function, the reasons that an individual has (or takes herself to have) for holding the attitude in question, and her level of self-awareness.

[27] Cf., e.g., Jurin & Fortner (2002).

[28] Cf., e.g., Weigel & Newman (1976).

[29] Cf. Scott & Willits (1994), Stern et *al.* (1995), Tarrant & Cordell (1997), Blake et *al.* (1997), Roberts & Bacon (1997), Schultz & Zelezny (1998), O'Connor et *al.* (1999), Olli et *al.* (2001), and Casey & Scott (2006).

[30] Cf. Vining & Ebreo (1992). A further study targeting recycling behavior is Ebreo et *al.* (1999).

[31] Schultz & Oskamp (1996). The third study reported in their contribution consists in a meta-analysis of 7 studies on the relationship between general environmental concern and observed recycling behavior.

is that the finding might be at least partly due to the specific composition of sample targeted by the study. For one thing, as we saw, the Value-Action gap tends to be larger for women than for men, and this fact – if it is a fact[32] – may go some way towards explaining why NEP scores failed to prove a good predictor of the targeted subjects' self-reported recycling behaviors. Moreover, the well-documented higher levels of environmental concern amongst young people may in the end reflect weakly held opinions that are hence less likely to influence both observed and (as in our case) self-reported behavior. As stated, however, both hypotheses stand clearly in need of further investigation.

Several limitations of the present pilot study suggest possible avenues for future research. The first thing to note is that the targeted sample was rather small. As a consequence, the absence of statistically significant effects in the treatment group may plausibly be attributed, at least in part, to the study's limited statistical power. Although directional patterns in line with its hypothesis were observed, a post hoc power analysis conducted using GPower[33] revealed that the study had only 43.27% power to detect a medium effect size (Cohen's $d = 0.40$) in a within subject design with 22 participants. This effect size threshold is not arbitrary, as recent meta-research suggests that Cohen's $d \approx 0.40$ represents the median effect size observed across a broad spectrum of psychological research.[34] Consequently, while the results of the present study cannot support the presence of a treatment effect, they also cannot conclusively rule one out. The relatively low achieved power underscores the need for larger samples in future work to draw more reliable conclusions about the potential cognitive impact of green nudges on self-attributed beliefs. Future research on the consequences of green nudging on student's environmental beliefs should hence ideally aim at larger sample sizes, which – in light of the above-mentioned severe difficulties with getting students to participate in in-person studies[35] – could perhaps be reached either by simply having more than one experimenter involved in the recruitment phase (as well as in the data gathering phase), or else by relying on an online-based experimental design, in which case one could think of an appropriate digital green nudge, such as an environmentally relevant default choice. A second limitation is represented by the unbalanced demographic composition of the targeted sample. For reasons already discussed, future research should hence aim at obtaining a more balanced women/men proportion in the sample. A third limitation of the study is that it mostly relied on a global attitude measure – i.e. the NEP Scale – in order to assess student's environmental beliefs. The only specific measure consisted in the two items specifically designed in order to elicit subjects' responses toward recycling. In hindsight, however, it would have perhaps been better to include in the self-report questionnaire more items targeting subjects' recycling behavior. A final limitation of the present study is represented by the apparent lack of a plausible explanation for the only significant main effect highlighted by the statistical analysis – i.e. higher NEP scores for the control, as opposed to the treatment group – which may indicate a weakness either in the overall design of the experiment or in the randomization procedure adopted.

The stated goal of the present chapter was to begin testing my self-interpretation model of nudged beliefs – i.e. start looking for empirical evidence for the hypothesis that, in virtue of SK-

---

[32] Cf. fn. 24 above.

[33] Faul et *al.* (2007).

[34] Cf. Funder & Ozer (2019).

[35] Cf. paragraph on "participants and design" in section 4.2 above.

related mechanisms, nudging individuals towards performing specific pro-environmental actions can have a positive impact on their self-reported environmental beliefs. Although such evidence has not been found (yet), I think that the above pilot study constitutes a valid first step toward exploring further this psychologically plausible, and yet so far largely neglected consequence of nudge-based policies. The study's main merit, I believe, consists in providing a useful illustration of how the abstract philosophical and psychological ideas and considerations out of which my theoretical proposal was originally developed can be operationalized, and therefore, at least in principle, empirically validated. As a consequence, or so I like to think, it constitutes a concrete starting point building on which one can design more elaborate experimental studies.

4.5 Appendix

Subject in the study were administered the following self-report questionnaire:[36]

QUESTIONARIO SELF-REPORT

Qui sotto trovi elencate delle affermazioni che riguardano il rapporto degli esseri umani con l'ambiente. Indica, per favore, il tuo grado di accordo con ognuna di esse su una scala da **1** (per nulla d'accordo) a **5** (estremamente d'accordo).

1. **Stiamo raggiungendo il limite massimo del numero di persone su questa Terra**

1	2	3	4	5

2. **Gli esseri umani hanno il diritto di modificare l'ambiente naturale per i propri bisogni**

1	2	3	4	5

3. **Quando gli esseri umani interferiscono con la natura, si producono effetti disastrosi**

1	2	3	4	5

4. **Grazie all'ingegno umano, la Terra rimarrà un luogo vivibile**

1	2	3	4	5

[36] The Italian translation is due to Gabriele Prati and Bruna Zani. Cf. Zani & Prati (2012).

5. **Gli esseri umani stanno abusando gravemente dell'ambiente**

1	2	3	4	5

6. **La Terra in realtà ha tante risorse naturali se solo sapessimo farne buon uso**

1	2	3	4	5

7. **Gli esseri umani hanno il dovere di tutelare la vita di animali e piante**

1	2	3	4	5

8. **L'equilibrio dell'ambiente è forte abbastanza da reggere l'impatto delle società industrializzate**

1	2	3	4	5

9. **Malgrado i progressi, siamo ancora in balia della forza della natura**

1	2	3	4	5

10. **I problemi ambientali sono stati in larga misura esagerati**

1	2	3	4	5

11. **La Terra ha risorse limitate**

1	2	3	4	5

12. **Gli esseri umani sono destinati a comandare sulla natura**

1	2	3	4	5

13. **L'equilibrio della natura è delicato e fragile**

1	2	3	4	5

14. Con il tempo gli esseri umani impareranno come funziona la natura e riusciranno a controllarla

1	2	3	4	5

15. Se le cose vanno avanti così, presto ci sarà una catastrofe ambientale

1	2	3	4	5

Qui sotto trovi due domande relative al tuo comportamento di riciclo. Indica, per favore, una risposta su una scala da **1** (mai) a **5** (sempre).

Quanto spesso fai la raccolta differenziata?

1	2	3	4	5

Quanto spesso fai attenzione all'imballaggio (riciclabile o non riciclabile) dei prodotti che acquisti?

1	2	3	4	5

The revised NEP Scale consists in the following 15 items:[37]

1. We are approaching the limit of the number of people the earth can support.
2. Humans have the right to modify the natural environment to suit their needs.
3. When humans interfere with nature it often produces disastrous consequences.
4. Human ingenuity will ensure that we do NOT make the earth unlivable.
5. Humans are severely abusing the environment.
6. The earth has plenty of natural resources if we just learn how to develop them.
7. Plants and animals have as much right as humans to exist.
8. The balance of nature is strong enough to cope with the impact of modern industrial nations.

[37] Cf. Dunlap et *al.* (2000: 433).

9. Despite our special abilities humans are still subject to the laws of nature.

10. The so-called "ecological crisis" facing humankind has been greatly exaggerated.

11. The earth is like a spaceship with very limited room and resources.

12. Humans were meant to rule over the rest of nature.

13. The balance of nature is very delicate and easily upset.

14. Humans will eventually learn enough about how nature works to be able to control it.

15. If things continue on their present course, we will soon experience a major ecological catastrophe.

Conclusions

Having an axe to grind, psychologists have long been warning all of us (and especially their fellow scientists), puts you in a tricky spot, as it can often make you see exactly what you want to see. This now well-documented phenomenon (confirmation bias), has however become so ingrained in our ordinary ways of thinking, that it sometimes prevents us from pausing to consider the other side of the epistemic medal: you can at times be right about something, even though you just happen to have a horse in the race. "Love it, or hate it" – writes the founder of the Behavioral Insight Team – "nudging is here to stay."[1]

As I hope to have shown in the present work, his understandable partiality to the matter notwithstanding, Halpern has a (huge) point. Regardless of how we feel about this wide-reaching social issue, over the last couple of decades (and arguably for much longer, albeit under different names) nudges have become a pervasive feature of our physical and digital environments. What this implies, I think, is that it is today in our best interest to seriously engage in the theoretical and experimental work required to reach a better grasp of the many ways in which nudge-based policies can and often do influence, for better or for worse, our decisions and choices – i.e. we need to open nudges' black box.

As I also hope to have shown, although nudging research is still in its infancy, the study of its psychology stands to gain a great deal from being seen and treated as the last chapter in a much longer story. A large chunk of that story, as we saw, is to be found in social psychology's long-standing interest in two interconnected issues. One consists in what Nisbett aptly calls 'the power of the situation' – i.e. the idea that the circumstances we find ourselves in (what in this work we referred to as our choice-architecture), whether we realize it or not (and we typically do not), have been amply shown to have a much larger impact on our thoughts and behaviors than we normally assume. The other issue concerns the many, and often counterintuitive ways in which our attitudes and behaviors are causally linked in reciprocal ways – i.e. the idea that, as it has likewise been abundantly established, our behaviors can influence our beliefs just as much (in fact, possibly more) as our beliefs can influence our behaviors.

To the extent that research on both of these fronts is still very much open and ongoing, joint efforts to develop a mature psychology of nudging can not only profitably build on the already existing, and hardly-acquired knowledge of the relevant phenomena – it can also validly contribute to the advancement of that knowledge.

An important lesson that emerges from a careful examination of this larger literature, as I tried to articulate throughout the present work, is that, to use a slogan, beliefs *matter* to nudging – i.e. the specific opinions that people happen to hold about various socially- and environmentally-relevant issues, not only have a sizable impact on the nudging process, but they are also likely, under certain conditions, to be affected by it. This, in my view, is where epistemological inquiry becomes directly relevant to nudging research. How do people normally come to know about what they think and feel about things? An answer well worth considering, I have suggested, is that acquiring Self-Knowledge often (and, in fact, more frequently than we realize) requires us to

[1] Cf. Halpern (2019: 12, 376).

become the observers of ourselves, and to engage in a mostly unconscious process of self-interpretation. As we saw, many today feel that – barring a few sound (yet largely unheeded) historical exceptions – most of the Western philosophical and psychological theorizing on the matter might have blown the powers of our introspection dramatically out of proportion. Evidence suggests, however, that – contrary to what their typical phenomenology would indeed suggest – most of our judgments about both our own mental states, and the mental states of others, crucially rely on evidence about our own and others past behaviors, as well as about the circumstances in which those behaviors occurred. In other words, such judgments seem to be largely inference-based, and not all a 'direct readout' of reality. In this regard – although this would be the topic for another, and hopefully future, work – my general sense is that under the pressure of an otherwise healthy cognitive revolution, and the consequent rush to systematically disavow the behaviorist tradition in all of its many different forms, several (likewise healthy) parts of the baby might have been hastily thrown away with the bath water. I suspect that I am not alone in feeling as I do, and I am also moderately confident that, provided that they steer clear of the 'surely operator', most philosophers will eventually come to acknowledge this point.

Going back to the present work, its main ambition has been to redress what I regard as an imbalance in current theoretical and empirical investigations of nudging – i.e. their nearly utter neglect of the possibility that nudge-based policies might indeed have an impact on people's beliefs, over and above their decisions and choices. Once one starts looking at the psychology of nudging as I (and others) suggested that we should – i.e. as just the last page in a much longer book – this hypothesis seems worthy of receiving more systematic attention than it has so far been the case. Nudging *is* indeed a powerful tool in the hands of natural born influencers. At the same time, however, it is definitely not this ingenious tool that first led them to try and influence each other's beliefs by manipulating their environment. The kind of activity that Levy refers to as *epistemic engineering* has indeed been part and parcel of our daily grind since the dawn of time. Today, we call it nudging, and we have all the means that we need to delve into a rigorous exploration it many possible consequences. Like it or not, each and every one of us needs to take some bets in life, come to terms with that great bookie that we like to call 'reality'. One of mines, is that a few years down the road other people will join me in thinking that the road I indicated in this dissertation – although currently among the 'less traveled by' – is one the taking of which could make 'all the difference'.[2]

[2] Frost (1949).

References

Abelson, R.P.A. E., McGuire, W.J., Newcomb, T.M., Rosenberg, M. J., Tannenbaum, R.H. (1968) *Theories of Cognitive Consistency : A sourcebook*. Chicago, IL: Rand McNally.

Albarracín, D., Wyer, R. (2000) The cognitive impact of past behavior: Influences on beliefs, attitudes, and future behavioral decisions. *Journal of Personality and Social Psychology* 79: 5 – 22. https://doi.org/10.1037//0022-3514.79.1.5

Allais, M. (1953) Le Comportement de l'Homme Rationnel Devant le Risque : Critiques des Postulats and Axiomes de l'Ecole Américaine. *Econometrica* 21: 503-46. https://doi.org/10.2307/1907921

Allcott, H. (2011) Social norms and energy conservation. *Journal of Public Economics* 95: 1082–95. https://doi.org/10.1016/j.jpubeco.2011.03.003

Allcott, H., Rogers, T. (2014) The short-run and long-run effects of behavioral interventions: experimental evidence from energy conservation. *American Economic Review* 104 (10): 3003–37. https://doi.org/10.1257/aer.104.10.3003

Alston, W. P. (1989) *Epistemic justification: Essays in the theory of knowledge*. Ithaca: Cornell University Press.

Ajzen, I. (1985) From intentions to actions: A theory of planned behavior. In J. Kuhl, J. Beckmann (eds.) *Action control: From cognition to behavior* (pp. 11–39). Berlin: Springer-Verlag.

Ajzen, I. (1991) The theory of planned behavior. *Organizational Behavior and Human Decision Processes* 50: 179-211. https://doi.org/10.4135/9781446249215.n22

Andersen, H. (2014) A field guide to mechanisms: Part I and II. *Philosophy Compass* 9(4): 274–293. https://doi.org/10.1111/phc3.12119 https://doi.org/10.1111/phc3.12118

Andrews, G. (1999) Efficacy, effectiveness and efficiency in mental health service delivery. *Australian and New Zealand Journal of Psychiatry* 33: 316–322. https://doi.org/10.1046/j.1440-1614.1999.00581.x

Andrews, A., Gregoire, M., Rasmussen, H., Witowich, G. (2013) Comparison of recycling outcomes in three types of recycling collection units. *Waste Management* 33: 530-35. http://dx.doi.org/10.1016/j.wasman.2012.08.018

Apperly, I. A. 2010. 2010. *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Abingdon: Psychology Press.

Arcury, T.A., Scollay, S.J., Johnson, T.P. (1987) Sex differences in environmental concern and knowledge: The case of acid rain. *Sex Roles* 16: 463-72. https://doi.org/10.1007/BF00292481

Ariely, D. (2008) *Predictably Irrational. The Hidden Forces that Shape Our Decisions*. London: HarperCollins.

Ariely, D., Norton, M.I. (2008) How actions create – not just reveal – preferences. *Trends in Cognitive Sciences* 12(1): 13-16. https://doi.org/10.1016/j.tics.2007.10.008

Armitage, C. J., Conner, M. (2001) Efficacy of the theory of planned behaviour: A meta-analytic review. *British Journal of Social Psychology* 40(4): 471- 499. https://doi.org/10.1348/014466601164939

Arno, A., & Thomas, S. (2016) The efficacy of nudge theory strategies in influencing adult dietary behaviour: A systematic review and meta-analysis. *BMC Public Health* 16(1): 1-11. https://doi.org/10.1186/s12889-016-3272-x

Aronson, E. (1969) The theory of cognitive dissonance: A current perspective. *Advances in Experimental Social Psychology* 4: 1–34. https://doi.org/10.1016/S0065-2601(08)60075-1

Aronson, E., Mills, J. M. (1959) The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology* 59: 177–181. https://doi.org/10.1037/h0047195

Aronson, E., Carlsmith, J. M. (1962) Performance expectancy as a determinant of actual performance. *Journal of Abnormal and Social Psychology* 65(3): 178-82. https://doi.org/10.1037/h0042291

Asch, S. E. (1951) Effects of Group Pressure on the Modification and Distortion of Judgments. In H. Guetzkow (ed.) *Groups, Leadership and Men*. Pittsburgh, PA: Carnegie Press. Pp. 177-90.

Asch, S. E. (1956) Studies of Independence and Conformity: I. A Minority of One Against a Unanimous Majority. *Psychological Monographs: General and Applied* 70(9): 1-70. https://doi.org/10.1037/h0093718

Axsom, D. (1989) Cognitive dissonance and behavior change in psychotherapy. *Journal of Experimental Social Psychology* 25: 234–252. https://doi.org/10.1016/0022-1031(89)90021-8

Bamberg, S. (2003) How does environmental concern influence specific environmentally related behaviors? A new answer to an old question. *Journal of Environmental Psychology* 23: 21–32. https://doi.org/10.1016/S0272-4944(02)00078-6

Bargh, J. A. (2006) *Social Psychology and the Unconscious*: *The Automaticity of Higher Mental Processes*. New York: Psychology Press.

Barkow, H., Cosmides, L., Tooby, J. (1992) *The Adapted Mind*. New York: Oxford University Press.

Baron-Cohen, S. (1997) *Mindblindness: An Essay on Autism and the Theory of Mind*. Cambridge, MA: MIT Press.

Barton, A., Grüne-Yanoff, T. (2015) From libertarian paternalism to nudging—And beyond. *Review of Philosophy and Psychology* 6(3): 341–359. https://doi.org/10.1007/s13164-015-0268-x

Beckenbach, F., Kahlenborn, W. (eds.) (2016) *New Perspectives for Environmental Policies Through Behavioral Economics*. Berlin: Springer. https://doi.org/10.1007/978-3-319-16793-0

Behavioural Insights Team (2020) *The little book of green nudges*. https://www.bi.team/publications/the-littlebook-of-green-nudges/

Bem, D.J. (1967) Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review* 74(3): 183-200. https://doi.org/10.1037/h0024835

Bem, D. J. (1972) Self-Perception Theory. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* 6: 1–62.

Benartzi, S., Beshears, J., Milkman, K.L., Sunstein, C.R., Thaler, R. H., Shankar, M., Tucker-Ray, W., Congdon, W.J., Galing, S. (2017) Should governments invest more in nudging? *Psychological Science* 28(8): 1041–1055. https://doi.org/10.1177/0956797617702501

Berg, N., Gigerenzer, G. (2010) As-if behavioral economics: Neoclassical economics in disguise? *History of Economic Ideas* 8(1): 133–165. https://doi.org/10.1400/140334

Bird, L., Wüstenhagen, R., Aabakken, J. (2002) A review of international green power markets: Recent experience, trends, and market drivers. *Renewable and Sustainable Energy Reviews* 6: 513–536. https://doi.org/10.1016/S1364-0321(02)00033-3

Blake, J. (1999) Overcoming the 'Value - action Gap' in Environmental Policy: Tensions between National Policy and Local Experience. *Local Environment* 4(3): 257–78. https://doi.org/10.1080/13549839908725599

Blocker, T. J., Eckberg, D.L. (1997) Gender and environmentalism: Results from the 1993 general social survey. *Social Science Quarterly* 78(4): 841-58.

Bohner, G., Schlüter, L.E. (2014) A Room with a Viewpoint Revisited: Descriptive Norms and Hotel Guests' Towel Reuse Behavior. PLoS One 9 (8): e104086. https://doi.org/10.1371/journal.pone.0106606

Bonini, N., Dorigoni, A. (2024) Green Nudging: A Behavioral Approach to Environmental Policies. In Singh, P., Daga, S. Yadav, K. (eds.) (2024) *Nudging Green: Behavioral Economics and Environmental Sustainability*. Cham: Springer Nature. Pp. 1 – 21.

Bord, R. J., Fisher, A., O'Connor, R.E. (1998) Public perceptions of global warming: United states and international perspectives. *Climate Research* 11: 75-84. https://doi.org/10.3354/cr011075

Bovens, L. (2009) The ethics of nudge. In *Preference change: Approaches from philosophy, economics and psychology*, eds. Till Grüne-Yanoff, and Sven Ove Hansson, 207-219. Berlin and New York: Springer Science & Business Media. https://doi.org/10.1007/978-90-481-2593-7_10

Boyd, R., Richerson, P. J., Henrich, J. (2011) The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences* 108 (Supplement 2): 10918-10925. https://doi.org/10.1073/pnas.1100290108

Breckler, S. J. (1984) Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of Personality and Social Psychology* 47(6): 1191-1205. https://doi.org/10.1037/0022-3514.47.6.1191

Brehm, J. W. (1956) Postdecision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology* 52: 384–9. https://doi.org/10.1037/h0041006

Brehm, J. W. (1966) *A theory of psychological reactance*. New York, NY: Academic Press.

Brehm, S. S., & Brehm, J. W. (1981) *Psychological reactance: A theory of freedom and control*. New York: Academic Press.

Briley, DA., Morris, M.W., Simonson, I. (2000) Reasons as Carriers of Culture: Dynamic versus Dispositional Models of Cultural Influence on Decision Making. *Journal of Consumer Research* 27(September):157–78. https://doi.org/10.1086/314318

Briñol, P., Petty, R. E. (2003) Overt head movements and persuasion: A self-validation analysis. *Journal of Personality and Social Psychology* 84: 1123–1139. https://doi.org/10.1037/0022-3514.84.6.1123

Bronchetti, E. T., Dee, T. S., Huffman, D. B., Magenheim, E. (2013) When a nudge isn't enough: Defaults and saving among low-income tax filers. *National Tax Journal* 66(3): 609–634. https://doi.org/10.17310/ntj.2013.3.04

Brown, C. L., Krishna, A. (2004) The skeptical shopper: A metacognitive account for effects of defaults options on choice. *Journal of Consumer Research* 31: 529–539. https://doi.org/10.1086/425087

Bucher, T., Collins, C., Rollo, M.E., McCaffrey, T.A., De Vlieger, N., Van Der Bend, D., Truby, H., Perez-Cueto, F.J.A. (2016) Nudging consumers towards healthier choices: a systematic review of positional influences on food choice. *British Journal of Nutrition* 115(12): 2252–63. https://doi.org/10.1017/S0007114516001653

Buchtel, E. E., Norenzayan, A. (2009) Thinking across cultures: Implications for dual processes. In Evans, St. B. T., and K. Frankish (eds.) *In two minds: Dual processes and beyond*, pp. 217-238, New York: Oxford University Press.

Budescu, D. V., Broomell, S., Por, H.-H. (2009) Improving communication of uncertainty in the reports of the Intergovernmental Panel on Climate Change. *Psychological Science* 20: 299–308. https://doi.org/10.1111/j.1467-9280.2009.02284.x

Calboli, S., Fano, V. (2022) Mechanistic explanations and the ethics of nudging. *Rivista Internazionale di Filosofia e Psicologia* 13: 175-186. https://doi.org/10.4453/rifp.2022.0017

Camerer, C., Issacharoff, G., Loewenstein, T., O'Donoghue, T., Rabin, M. (2003) Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism'. *University of Pennsylvania Law Review* 1151: 1211-54. https://doi.org/10.2307/3312889

Campbell, D. T. (1963) Social attitudes and other acquired behavioral dispositions. In K. Sigmund (ed.) *Psychology: A study of a science. Study II. Empirical substructure and relations with other sciences. Investigations of man as socius: Their place in psychology and the social sciences* (Vol. 6, pp. 94–172). New York: McGraw-Hill.

Carlsmith, J. M., Collins, B. E., Helmreich, R. L. (1966) Studies in forced compliance: I. The effect of pressure for compliance on attitude change produced by face-to-face role playing and anonymous essay writing. *Journal of Personality and Social Psychology* 4(1): 1-13. https://doi.org/10.1037/h0023507

Carlsson, F., Gravert, C., Johansson-Stenman, O., Kurz, V. (2019) Nudging as an Environmental Policy Instrument. CeCAR Working Paper Series No. 4. http://doi.org/10.2139/ssrn.3711946

Carruthers, P. (2011) *The Opacity of Mind. An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Cartwright, N. (2009) What is this thing called 'efficacy'? In C. Mantzavinos (ed.) *Philosophy of the Social Sciences. Philosophical Theory and Scientific Practice*, pp. 185–206. Cambridge: Cambridge University Press.

Cartwright, N., Hardie, J. (2012) *Evidence-based Policy: A Practical Guide to Doing it Better*. New York: Oxford University Press.

Casey, P. J., Scott, K. (2006) Environmental concern and behavior in an Australian sample within an ecocentric-anthropocentric framework. *Australian Journal of Psychology* 58(2): 57-67. https://doi.org/10.1080/00049530600730419

Cassam, Q. (2015) *Self-Knowledge for Humans*, Oxford: Oxford University Press.

Chaiken, S. (1980) Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology* 39(5): 752–766. https://doi.org/10.1037/0022-3514.39.5.752

Chaiken, S. (1987) The heuristic model of persuasion. In M. P. Zanna, J. M. Olson, & C. P. Herman (Eds.), *Social influence: The Ontario symposium* (Vol. 5, pp. 3–39). Mahwah, NJ: Lawrence Erlbaum Associates.

Chaiken, S., Baldwin, M. W. (1981) Affective-Cognitive Consistency and the Effect of Salient Behavioral Information on the Self-Perception of Attitudes. *Journal of Personality and Social Psychology* 41 (1): 1 – 12. https://doi.org/10.1037/0022-3514.41.1.1

Chaiken, S., Pomerantz, E. M., Giner-Sorolla, R. (1995) Structural consistency and attitude strength. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences*. Ohio State University series on attitudes and persuasion (Vol. 4., pp. 387–412). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chaiken, S., Trope, Y. (eds.) (1999) *Dual-process theories in social psychology*. New York: Guilford Press.

Chaiken, S., Ledgerwood, A. (2012) A theory of heuristic and systematic information processing. In P.A.M. van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 246–266). Thousand Oaks, CA: Sage Publications.

Chater, N. (2018) *The Mind is Flat. The Illusion of Mental Depth and the Improvised Mind*. Penguin Books.

Cialdini, R. B. (2021) *Influence. The psychology of persuasion*. HarperCollins.

Cialdini, R.B., Reno, R.R., Kallgren, C.A. (1990) A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology* 58(6): 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Clark, C. F., Kotchen, M. J., Moore, M. R. (2003) Internal and external influences on pro-environmental behavior: Participation in a green electricity program. *Journal of Environmental Psychology* 23: 237–246. https://doi.org/10.1016/S0272-4944(02)00105-6

Congiu, L., Moscati, I. (2021) A review of nudges: Definitions, justifications, effectiveness. *Journal of Economic Surveys* 36(1): 188-213. https://doi.org/10.1111/joes.12453

Costa, D.L., Kahn, M.E. (2013) Energy conservation "nudge" and environmentalist ideology: evidence from a randomized residential field experiment. *Journal of the European Economic Association* 11 (3): 680–702. https://doi.org/10.1111/jeea.12011

Craver, C. F. (2006) What mechanistic models explain. *Synthese* 153: 355–376. https://doi.org/10.1007/s11229-006-9097-x

Craver, C., Tabery, J., Illari, P. (2024) Mechanisms in Science. In Edward N. Zalta & Uri Nodelman (eds.) *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition). URL = <https://plato.stanford.edu/archives/fall2024/entries/science-mechanisms/>.

Congiu, L., Moscati, I. (2022) A review of nudges: Definitions, justifications, effectiveness. *Journal of Economic Surveys* 36(1): 188–213. https://doi.org/10.1111/joes.12453

Cook, J., Oreskes, N., Doran, P.T., Anderegg, W.R.L., Verheggen, B., Maibach, E.W., Stuart Carlton, J. et *al.* (2016) Consensus on consensus: a synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters* 11(4): 048002. https://doi.org/10.1088/1748-9326/11/4/048002

Cooke, R., Sheeran, P. (2004) Moderation of cognition-intention and cognition-behaviour relations: A meta-analysis of properties of variables from the theory of planned behaviour. *British Journal of Social Psychology* 43(2): 159–186. https://doi.org/10.1348/0144666041501688

Cooper, J. (2007) *Cognitive Dissonance: 50 Years of a Classic Theory*. SAGE Publications.

Cooper, J., Worchel, S. (1970) Role of undesired consequences in arousing cognitive dissonance. *Journal of Personality and Social Psychology* 16(2): 199–206. https://doi.org/10.1037/h0029830

Cooper, J., Fazio, R. H. (1984) A new look at dissonance theory. *Advances in Experimental Social Psychology* 17: 229–266. https://doi.org/10.1016/S0065-2601(08)60121-5

Cooper, J., Blackman, S.F., Keller, K. T. (2016) *The Science of Attitudes*. New York: Routledge.

Cosic A., Cosic H., Ille S. (2018) Can nudges affect students' green behaviour? A field experiment. *Journal of Behavioral Economics for Policy* 2(1): 107-111.

Costa, D.L., Kahn, M.E. (2013) Energy conservation "nudge" and environmentalist ideology: evidence from a randomized residential field experiment. *Journal of the European Economic Association* 11(3): 680–702. https://doi.org/10.1111/jeea.12011

Croteau J. (2019), *Mind the gap: the value-action gap, nudges, and an ecosocial vision*. Fort Collins, CO: Department of Political Science, Colorado State University. Thesis.

Davidson, D.J., Freudenburg, W.R. (1996) Gender and environmental risk concerns: A review an analysis of available research. *Environment and Behavior* 28(3): 302-39. https://doi.org/10.1177/0013916596283003

Dayan, E., Bar-Hillel, M. (2011) Nudge to nobesity II: Menu positions influence food orders. *Judgment and Decision Making* 6(4): 333–342. https://doi.org/10.1017/S1930297500001947

Dennett, D. C. (2013) *Intuition pumps and other tools for thinking*. Norton & Company.

Dewies, M., Schop-Etman, A., Rohde, K.I.M., Denktaş, S. (2021) Nuding is Ineffective When Attitudes Are Unsupportive: An Example from a Natural Field Experiment. *Basic and Applied Social Psychology* 43(4): 213-25. https://doi.org/10.1080/01973533.2021.1917412

Diez, T., Kalof, L., Stern, P.C. (2002) Gender, values, and environmentalism. *Social Science Quarterly* 83(1): 353-64. https://doi.org/10.1111/1540-6237.00088

Dogramaci, S. (2012) Reverse Engineering Epistemic Evaluations. *Philosophy and Phenomenological Research* 84(3): 513-530. https://doi.org/10.1111/j.1933-1592.2011.00566.x

Doob, L. W. (1947) The behavior of attitudes. *Psychological Review* 54(3): 135-156. https://doi.org/10.1037/h0058371

Dunlap, R. E., Van Liere, K. D. (1978) The 'new environmental paradigm'. *The Journal of Environmental Education* 9(4): 10–19. https://doi.org/10.1080/00958964.1978.10801875

Dunlap, R. E., Van Liere, K. D., Mertig, A. G., Jones, R. E. (2000) New trends in measuring environmental attitudes: Measuring endorsement of the new ecological paradigm: A revised NEP scale. *Journal of Social Issues* 56(3): 425–442. https://doi.org/10.1111/0022-4537.00176

Dunlap, R. E. (2008) The New Environmental Paradigm Scale: From Marginality to Worldwide Use. *The Journal of Environmental Education* 40(1): 3-18. https://doi.org/10.3200/JOEE.40.1.3-18

Ebreo, A., Hershey, J., Vining, J. (1999) Reducing solid waste. Linking recycling to environmentally responsible consumerism. *Environment and Behavior* 31(1): 107–135. https://doi.org/10.1177/00139169921972029

Elberg Nielsen, A.S., Sand, H., Sørensen, P., Knutsson, M., Martinsson, P., Persson, E., Wollbrant, C. (2016) *Nudging and pro-environmental behaviour.* Nordic Council of Ministers: Rosendahls, DK.

Ellsberg, D. (1961) Risk, Ambiguity and the Savage Axioms. *Quarterly Journal of Economics* 75: 643-69. https://doi.org/10.2307/1884324

Emens, E. (2007) Changing Name Changing: Framing Rules and the Future of Marital Names. *University of Chicago Law Review* 74(3): 761–863.

Engelen, B. (2019) Nudging and rationality: What is there to worry? *Rationality and Society* 31(2): 204–232. https://doi.org/10.1177/1043463119846743

Evans, St. B. T. (2008) Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology* 59: 255-78. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, St. B. T., Barston, J.L., Pollard, P. (1983) On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition* 11: 295-306. https://doi.org/10.3758/BF03196976

Evans, St. B. T., Over, D. E. (1996) *Rationality and reasoning*. Hove: Psychology Press.

Evans, St. B. T., Frankish, K. (eds.) (2009) *In two minds: Dual processes and beyond*. New York: Oxford University Press.

Evans, N., Eickers, S., Geene, L., Todorovic, M., Villmow, A. (2017) Green Nudging. A discussion and preliminary evaluation of nudging as an environmental policy instrument. FFU-Report 01-2017.

Fahy, F. (2005) The Right to Refuse: Public Attitudes and Behaviour towards Waste in West Ireland. *Local Environment* 10(6): 551-569. https://doi.org/10.1080/13549830500321618

Farrow, K, Grolleau, G., Ibanez, L. (2017) Social Norms and Pro-environmental Behavior: A Review of the Evidence. *Ecological Economics* 140: 1-13. https://doi.org/10.1016/j.ecolecon.2017.04.017

Faul, F., Erdfelder, E., Lang, A.G., Buchner, A. (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39(2): 175-191. https://doi.org/10.3758/bf03193146

Fazio, R. H. (1986) How do attitudes guide behavior. *Handbook of motivation and cognition: Foundations of social behavior* 1: 204–243.

Fazio, R. H. (1990) Multiple Processes by which Attitudes Guide Behavior: The Mode Model as an Integrative Framework. *Advances in Experimental Social Psychology* 23: 75–109. https://doi.org/10.1016/S0065-2601(08)60318-4

Fazio, R. H., & Towles-Schwen, T. (1999) The MODE model of attitude-behavior processes. In S. Chaiken, and Y. Trope (eds.) *Dual process theories in social psychology* (pp. 97–116). New York: Guilford Press.

Festinger, L. (1957) *A Theory of Cognitive Dissonance*. Stanford University Press.

Festinger, L., & Macoby, N. (1964). On resistance to persuasive communication. *Journal of Abnormal and Social Psychology* 68(4): 359–366. https://doi.org/10.1037/h0049073

Fishbein, M. (2000) The role of theory in HIV prevention. *AIDS Care* 12(3): 273–278. https://doi.org/10.1080/09540120050042918

Fishbein, M., Ajzen, I. (1974) Attitudes towards objects as predictors of single and multiple behavioral criteria. *Psychological Review* 81(1): 59-74. https://doi.org/10.1037/h0035872

Fishbein, M., Ajzen, I. (1975) *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Fishbein, M., Ajzen, I. (2010) *Predicting and changing behavior*. New York: Psychology Press.

Flynn, R., Bellaby, P., Ricci, M. (2009) The 'Value-Action Gap' in Public Attitudes towards Sustainable Energy: The Case of Hydrogen Energy. *The Sociological Review* 57(2_suppl): 159-180. https://doi.org/10.1111/j.1467-954X.2010.01891.x

Frank, R. H. (2020) *Under the influence. Putting peer pressure to work*. Princeton University Press.

Frankfurt, H. G. (1971) Freedom of the will and the concept of a person. *The Journal of Philosophy* 68(1): 5–20. https://doi.org/10.2307/2024717

Frankish, K., Evans, St. B. T. (2009) The duality of mind: An historical perspective. In Evans, St. B. T., and K. Frankish (eds.) *In two minds: Dual processes and beyond*, pp. 1-29, New York: Oxford University Press.

Fricker, E. (2006) Testimony and Epistemic Autonomy. In *The Epistemology of Testimony*, eds. Jennifer Lackey, Ernest Sosa, 225-251. New York: Oxford University Press.

Friedman, M., Savage, L. J. (1948) The Utility of Choices Involving Risk. *Journal of Political Economy* 56: 279-304. https://doi.org/10.1086/256692

Frigg, R., Hartmann, S. (2024) Models in Science. In Edward N. Zalta & Uri Nodelman (eds.) *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition). URL = <https://plato.stanford.edu/archives/fall2024/entries/models-science/>.

Froming, W. J., Walker, G. R., Lopyan, K. J. (1982) Public and private self-awareness: When personal attitudes conflict with societal expectations. *Journal of Experimental Social Psychology* 18(5): 476–487. https://doi.org/10.1016/0022-1031(82)90067-1

Frost, R. (1949) *Complete Poems of Robert Frost*. New York: Henry Holt and Company.

Funder, D.C., Ozer, D.J. (2019) Evaluating effect size in psychological research: sense and nonsense. *Advances in Methods and Practices in Psychological Science* 2(2): 156-168. https://doi.org/10.1177/2515245919847202

Gawronski, B., Strack, F. (eds.) (2012) *Cognitive consistency: A fundamental principle in social cognition*. New York: Guilford Press.

Gertler, B., (2024) Self-Knowledge. In E. N. Zalta, U. Nodelman (eds.) *The Stanford Encyclopedia of Philosophy* (Summer 2024 Edition). URL=<https://plato.stanford.edu/archives/sum2024/entries/self-knowledge/>.

Gifford. R. (2011) The dragons of inaction: Psychological barriers that limit climate change mitigation and adaptation. *American Psychologist* 66 (4): 290 – 302. https://doi.org/10.1037/a0023566

Gifford. R. (2014) Environmental Psychology Matters. *Annual Review of Psychology* 65: 541-79. https://doi.org/10.1146/annurev-psych-010213-115048

Gifford, R., Scannell, L., Kormos, et al. (2009) Temporal pessimism and spatial optimism in environmental assessments: an 18-nation study. *Journal of Environmental Psychology* 29 (1):1-12. https://doi.org/10.1016/j.jenvp.2008.06.001

Gifford, R., Kormos, C., McIntyre, A. (2011) Behavioral dimensions of climate change: drivers, responses, barriers, and interventions. *WIREs Climate Change* 2: 801–827. https://doi.org/10.1002/wcc.143

Gifford, R. Comeau, L. A. (2011) Message framing influences perceives climate change competence, engagement, and behavioral intentions. *Global Environmental Change* 21: 1301-1307. https://doi.org/10.1016/j.gloenvcha.2011.06.004

Gifford, R., Sussman, R. (2012) Environmental attitudes. In S. D. Clayton (Ed.), *The Oxford handbook of environmental and conservation psychology*. Oxford University Press, pp. 65 – 80.

Gifford, R., Nilsson, A. (2014) Personal and social factors that influence pro-environmental concern and behaviour: A review. *International Journal of Psychology* 49(3): 141-157. https://doi.org/10.1002/ijop.12034

Goethals, G. R., Cooper, J., Naficy, A. (1979) Role of foreseen, foreseeable, and unforeseeable behavioral consequences in the arousal of cognitive dissonance. *Journal of Personality and Social Psychology* 37(7): 1179–85. https://doi.org/10.1037/0022-3514.37.7.1179

Goldstein, N.J., Cialdini, R.B., Griskevicius, V. (2008) A room with a viewpoint: using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research* 35(3): 472–82. https://doi.org/10.1086/586910

Gopnik, A. (1993) How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, 16(1): 1–14. https://doi.org/10.1017/S0140525X00028636

Graziano, M., Gillingham, K. (2015) Spatial patterns of solar photovoltaic system adoption: The influence of neighbors and the built environments. *Journal of Economic Geography* 15: 815-839. https://doi.org/10.1093/jeg/lbu036

Greenwald, A. G. (1968) Cognitive learning, cognitive response to persuasion, and attitude change. In G. Greenwald, T. C. Brock, & T. M. Ostrom (Eds.), *Psychological foundations of attitudes* (pp. 147–170). New York: Academic Press.

Gromet, D.M., Kunreuther, H., Larrick, R.P. (2013) Political identity affects energy efficiency attitudes and choices. *Proceedings of the National Academy of Sciences* 110 (23): 9314–9319. https://doi.org/10.1073/pnas.1218453110

Grundmann, T. (2021) The Possibility of Epistemic Nudging. *Social Epistemology* 37(2): 208-18. https://doi.org/10.1080/02691728.2021.1945160

Grüne-Yanoff, T. (2012) Old wine in new casks: Libertarian paternalism still violates liberal principles. *Social Choice and Welfare* 38(4): 635–645. https://doi.org/10.1007/s00355-011-0636-0

Grüne-Yanoff, T. (2016) Why behavioural policy needs mechanistic evidence. *Economics and Philosophy* 32 (3): 463–483. https://doi.org/10.1017/S0266267115000425

Halpern, D. (2019) *Inside the Nudge Unit: How Small Changes Can Make a Big Difference*. London: Penguin.

Hansen, P. G. (2019) Nudging: To know 'what works' you need to know why it works. *Journal of Behavioral Economics for Policy* 3(Special Issue): 9-11.

Hansen, P. G., Jespersen, A. M. (2013) Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation* 4(3): 3–28. https://doi.org/10.1017/S1867299X00002762

Harmon-Jones, E., Harmon-Jones, C. (2002) Testing the action-based model of cognitive dissonance: The effect of action orientation on postdecisional attitudes. *Personality and Social Psychology Bulletin* 28(6): 711–723. https://doi.org/10.1177/0146167202289001

Hauser, O.P., Gino, F., Norton, M.I. (2018) Budging beliefs, nudging behavior. Mind & Society 17: 15-26. https://doi.org/10.1007/s11299-019-00200-9

Hausman, D.M., Welch, B. (2010) Debate: To nudge or not to nudge. *Journal of Political Philosophy* 18(1): 123–136. https://doi.org/10.1111/j.1467-9760.2009.00351.x

Hawcroft, L. J., Milfont, T. L. (2010) The use (and abuse) of the new environmental paradigm scale over the last 30 years: A meta-analysis. *Journal of Environmental Psychology* 30(2): 143–158. https://doi.org/10.1016/j.jenvp.2009.10.00

Hine, D. W., Gifford, R. (1996). Individual restraint and group efficiency in commons dilemmas: The effects of two types of environmental uncertainty. *Journal of Applied Social Psychology* 26 11): 993–1009. https://doi.org/10.1111/j.1559-1816.1996.tb01121.x

Honnold, J. A. (1984) Age and environmental concern: some specification of effects. *Journal of Environmental Education* 16(1): 4-9. https://doi.org/10.1080/00958964.1984.9942695

Horowitz, Sophie, "Higher-Order Evidence", *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/fall2022/entries/higher-order-evidence/

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953) *Communication and persuasion: Psychological studies of opinion change*. New Haven, CT: Yale University Press.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. (1949) *Experiments on mass communication*. Princeton, NJ: Princeton University Press.

Hovland, C. I., Mandell, W. (1952) An experimental comparison of conclusion-drawing by the communicator and by the audience. *The Journal of Abnormal and Social Psychology*

47(3): 581–588. https://doi.org/10.1037/h0059833

Hovland, C. I., Weiss, W. (1951) The influence of source credibility on communication effectiveness. *Public Opinion Quarterly* 15(4): 635–650. https://doi.org/10.1086/266350

Hovland, C. I., Harvey, O. J., Sherif, M. (1957) Assimilation and contrast in communication and attitude change. *Journal of Abnormal and Social Psychology* 55(2): 242–252. https://doi.org/10.1037/h0048480

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory*. NY: Appelton-Century-Crofts.

Humphrey, N. (1984) *Consciousness Regained. Chapters in the Development of Mind*. Oxford: Oxford University Press.

Illari, P. M., Williamson, J. (2012) What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science* 2: 119–135. https://doi.org/10.1007/s13194-011-0038-2

IPCC, Climate Change 2021 – The Physical Science Basis Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. https://www.ipcc.ch/report/ar6/wg1/

Janis, I. L., Feshbach, S. (1953) Effects of fear-arousing communications. *Journal of Abnormal and Social Psychology* 48(1): 78–92. https://doi.org/10.1037/h0060732

Janis, I. L., Gilmore, J. B. (1965) The influence of incentive conditions on the success of role playing in modifying attitudes. *Journal of Personality and Social Psychology* 1(1): 17-27. https://doi.org/10.1037/h0021643

Johnson, E. J., Goldstein, D. (2003) Do defaults save lives? *Science* 302(5649): 1338–1339. https://doi.org/10.1126/science.1091721

Jones, R., Pykett, J., Whitehead, M. (2011) Governing temptation: Changing behaviour in an age of libertarian paternalism'. *Progress in Human Geography* 35 (4): 483–501. http://dx.doi.org.ezproxy.princeton.edu/10.1177/0309132510385741.

Jurin, R.R., Fortner, R.W. (2002) Symbolic beliefs as barriers to responsible environmental behavior. *Environmental Education Research* 8(4): 373-94. https://doi.org/10.1080/1350462022000026791

Just, D., Wansink, B. (2009) Smarter Lunchrooms: Using Behavioural Economics to Improve Meal Selection. *Choices* 24(3), Retrieved from: http://www.choicesmagazine.org/UserFiles/file/article_87.pdf.

Kahneman, D. (2011) *Thinking, fast and slow*. Farrar, Straus and Giroux, New York.

Kahneman, D., Slovic, P., Tversky, A. (eds.) (1982) *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kahneman, D., Frederich, S. (2002) Representativeness revisited: attribute substitution in intuitive judgment. In *Heuristics and Biases: The Psychology of Intuitive Judgment*, ed. T. Gilovich, D. Griffin, D. Kahneman, pp. 49-81. Cambridge, UK: Cambridge University Press.

Kahneman, D., Frederich, S. (2005) A model of heuristic judgment. In *The Cambridge Handbook of Thinking and Reasoning*, ed. K. Holyoak, R.G. Morrison, pp. 267-94. Cambridge, UK: Cambridge University Press.

Kelman, H. C., Hovland, C. I. (1953) "Reinstatement" of the communicator in delayed measurement of opinion change. *Journal of Abnormal and Social Psychology* 48(3): 327–335. https://doi.org/10.1037/h0061861

Kihlstrom, J. F. (1987) The cognitive unconscious. *Science* 237(4821): 1445-52. https://doi.org/10.1126/science.3629249

Klauer, K.C., Musch, J., Naumer, B. (2000) On belief bias in syllogistic reasoning. *Psychological Review* 107(4): 852 -84. https://doi.org/10.1037/0033-295X.107.4.852

Klein, G. (1999) *Sources of Power*. Cambridge, MA: MIT Press.

Kollmuss, A., Agyeman, J. (2002) Mind the Gap: Why do people act environmentally and what are the barriers to pro-environmental behavior? *Environmental Education Research* 8(3): 239 – 260. https://doi.org/10.1080/13504620220145401

Kraus, S. J. (1995) Attitudes and the prediction of behavior: A meta-analysis of the empirical literature. *Personality and Social Psychology Bulletin* 21(1): 58–75. https://doi.org/10.1177/0146167295211007

Kruglanski, A. W., Jasko, K., Milyavsky, M., Chernikova, M., Webber, D., Pierro, A., di Santo, D. (2018) Cognitive Consistency Theory in Social Psychology: A Paradigm Reconsidered. *Psychological Inquiry* 29(2): 45-59. https://doi.org/10.1080/1047840X.2018.1480619

Kuyer, P., Gordijn, B. (2023) Nudge in perspective: A systematic literature review on the ethical issues with nudging. *Rationality and Society* 35(2): 191-230. https://doi.org/10.1177/10434631231155005

LaPiere, R. T. (1934) Attitudes vs. actions. *Social Forces* 13(2): 230–237. https://doi.org/10.2307/2570339

Lacroix, K., Gifford, R. (2017) Psychological Barriers to Energy Conservation Behavior: The Role of Worldviews and Climate Change Risk Perception. *Environment and Behavior* 50 (7): 749–780. https://doi.org/10.1177/0013916517715296

Lasswell, H. D. (1948). The structure and function of communication in society. In L. Bryson (Ed.), *The communication of ideas: Religion and civilization series* (pp. 37–51). New York: Harper & Row.

Lehner, M., Mont, O., Heiskanen, E. (2016) Nudging— a promising tool for sustainable consumption behavior? *Journal of Cleaner Production* 134: 166-77. https://doi.org/10.1016/j.jclepro.2015.11.086

Levav, J., Fitzsimons, G. J. (2006) When questions change behavior: The role of ease of representation. *Psychological Science* 17(3): 207–213. https://doi.org/10.1111/j.1467-9280.2006.01687.x

Levy, Neil. 2022. *Bad Beliefs. Why They Happen to Good People*. Oxford: Oxford University Press.

Linder, D. E., Cooper, J., Jones, E. E. (1967) Decision freedom as a determinant of the role of incentive magnitude in attitude change. *Journal of Personality and Social Psychology* 6(3): 245-54. https://doi.org/10.1037/h0021220

Löfgren, Å., Martinsson, P., Hennlock, M., Sterner, T. (2012) Are experienced people affected by a pre-set default option – Results from a field experiment. *Journal of Environmental Economics and Management* 63: 66–72. http://doi.org/10.1016/j.jeem.2011.06.002

Lorenzoni, I., Nicholson-Cole, S., Whitmarsh, L. (2007) Barriers perceived to engaging with climate change among the UK public and their policy implications. *Global Environmental Change* 17 (3-4):445-59. https://doi.org/10.1016/j.gloenvcha.2007.01.004

MacKay, D., Robinson, A. (2016) The ethics of organ donor registration policies: Nudges and respect for autonomy. *The American Journal of Bioethics* 16(11): 3–12. https://doi.org/10.1080/15265161.2016.1222007

Marchionni, C., Reijula, S. (2019) What is mechanistic evidence, and why do we need it for evidence-based policy? *Studies in History and Philosophy of Science* 73: 54–63. https://doi.org/10.1016/j.shpsa.2018.08.003

Marchiori, D. R., Adriaanse, M. A., de Ridder, D. T. (2017) Unresolved questions in nudging research: Putting the psychology back in nudging. *Social and Personality Psychology Compass* 11(1): e12297. https://doi.org/10.1111/spc3.12297

Matheson, Jonathan. 2024. *Why It's OK Not to Think for Yourself*. New York, London: Routledge.

McCright, A.M., Dunlap, R.E. (2011) The politicization of climate change and polarization in the American's public view of global warming, 2001-2010. *The Sociological Quarterly* 52(2): 155–194. https://doi.org/10.1111/j.1533-8525.2011.01198.x

McGuire, W. J. (1961) Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *Journal of Abnormal and Social Psychology* 63(2): 326–332. https://doi.org/10.1037/h0048344

McKenzie, C.R.M. (2004) Framing effects in inference tasks—and why they are normatively defensible. *Memory & Cognition* 32: 874–885. https://doi.org/10.3758/BF03196866

McKenzie, C.R.M., Liersch, M.J., Finkelstein, S.R. (2006) Recommendations Implicit in Policy Defaults. *Psychological Science* 17 (5): 414-420. https://doi.org/10.1111/j.1467-9280.2006.01721.x

Mercier, H., Sperber, D. (2017) *The Enigma of Reason: A New Theory of Human Understanding*. London: Allen Lane.

Mercier, H. (2020) *Not Born Yesterday: The Science of Who We Trust and What We Believe*. Princeton, NJ: Princeton University Press.

Milfont, T. L. (2008) The effects of social desirability on self-reported environmental attitudes and ecological behavior. *The Environmentalist* 29: 263-69. https://doi.org/10.1007/s10669-008-9192-2

Miller, N. D., Meindl, J. N., Caradine, M. (2016) The effects of bin proximity and visual prompts on recycling in a university building. *Behavior and Social Issues* 25: 4-10. https://doi.org/10.5210/bsi.v25i0.6141

Mitchell, T. R., Thompson, L., Peterson, E., Cronk, R. (1997) Temporal adjustments in the evaluation of events: The "rosy view". *Journal of experimental social psychology* 33(4): 421-448. https://doi.org/10.1006/jesp.1997.1333

Mols, F., Haslam, A., Jetten, J., Steffens, N.K. (2015) Why a nudge is not enough: a social identity critique of governance by stealth. *European Journal of Political Research* 54(1): 81–98. https://doi.org/10.1111/1475-6765.12073

Moratti, A. (2020) Tecniche di nudging in ambito ambientale. Una rassegna di esperienze e risultati. Milano: Fondazione Cariplo.

Nagatsu, M. (2015) Social nudges: Their Mechanisms and Justification. *Review of Philosophy and Psychology* 6(3): 481–494. https://doi.org/10.1007/s13164-015-0245-4

Nisbett, R. E. (2016) *Mindware. Tools for Smart Thinking*. London: Penguin Books.

Nisbett, R. E., Wilson, T. D (1977) Telling more than we can know: verbal reports on mental processes. *Psychological Review* 84(3): 231-95. https://doi.org/10.1037/0033-295X.84.3.231

Nisbett, R. E., Ross, L. (1980) *Human inference. Strategies and Shortcomings of Social Judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Nolan, J.M., Wesley Schultz, P., Cialdini, R.B., Goldstein, N., Griskevicius, V. (2008) Normative social influence is underdetected. *Personality and Social Psychology Bulletin* 34(7): 913–923.

O'Connor, C., Goldberg, S., Goldman, A. (2014) Social Epistemology. In E. N. Zalta, U. Nodelman (eds.) *The Stanford Encyclopedia of Philosophy* (Summer 2024 Edition). URL = <https://plato.stanford.edu/archives/sum2024/entries/epistemology-social/>.

O'Connor, R. T., Lerman, D. C., Fritz, J. N., Hodde, H. B. (2010) Effects of number and location of bins on plastic recycling at a university. *Journal of Applied Behavior Analysis* 43(4): 711–715. https://doi.org/10.1901/jaba.2010.43-711

Oliver, A. (2013) From nudging to budging: Using behavioural economics to inform public sector policy. *Journal of Social Policy* 42(4), 685–700. https://doi.org/10.1017/S0047279413000299

Oliver, A. (2017) *The origins of behavioural public policy*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108225120

Ölander, F., Thøgersen, J. (2014) Informing versus nudging in environmental policy. *Journal of Consumer Policy* 37: 341–356. https://doi.org/10.1007/s10603-014-9256-2

Olli, E., Grendstad, G., Wollebaek, D. (2001) Correlates of environmental behaviors: Bringing back social context. *Environment and Behavior* 33(2): 181‑208. https://doi.org/10.1177/0013916501332002

Oskamp, S., Harrington, M. J., Edwards, T. C., Sherwood, D. L., Okuda, S. M., Swanson, D. C. (1991) Factors influencing household recycling behavior. *Environment and Behavior* 23(4): 494–519. https://doi.org/10.1177/0013916591234005

Osman, M., McLachlan, S., Fenton, N., Neil, M., Löfstedt, R., Meder, B. (2020) Learning from behavioural changes that fail. *Trends in Cognitive Sciences* 24(12): 969–980. https://doi.org/10.1016/j.tics.2020.09.009

Osterhouse, R. A., Brock, T. C. (1970) Distraction increases yielding to propaganda by inhibiting counterarguing. *Journal of Personality and Social Psychology* 15(4): 344–58. https://doi.org/10.1037/h0029598

Pavlov, I. (1927). *Conditioned reflexes*. Oxford, UK: Oxford University Press.

Pettit, P. (2014) *Just freedom: A moral compass for a complex world*. New York: W. W. Norton & Company.

Petty, R. E., Brock, T. C. (1981) Thought disruption and persuasion: Assessing the

validity of attitude change experiments. In Petty, R. E., Ostrom, T.M., Brock. T.C. (eds) *Cognitive responses in persuasion* (pp. 55–79). Hillsdale, NJ: Erlbaum.

Petty, R. E., Cacioppo, J. T. (1981) Issue involvement as a moderator of the effects on attitude of advertising content and context. *Advances in Consumer Research* 8(1): 20–24.

Petty, R. E., Cacioppo, J. T. (1984) The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology* 46(1): 69–81. https://doi.org/10.1037/0022-3514.46.1.69

Petty, R. E., Briñol, P. (2012) The Elaboration Likelihood Model. In P.A.M. Van Lange, A. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (Vol.1, pp. 224–245). London: Sage Publications.

Petty, R. E., Wheeler, S.C., Tomala, Z.L. (2003) Persuasion and attitude change. In T. Millon and M.J. Lerner (eds) *Handbook of Psychology: Personality and social psychology*, vol. 5, 353-82. John Wiley and Sons, Hoboken, NJ.

Pichert, Daniel, and Konstantinos V. Katsikopoulos. 2008. Green defaults: Information presentation and pro-environmental behavior. *Journal of Environmental Psychology* 28: 63–73. https://doi.org/10.1016/j.jenvp.2007.09.004

Pinker, Steven. 1994. *The Language Instinct*. London: Penguin.

Pirni, A. (2023) Climate Change and the Motivational Gap. In *Handbook of Philosophy of Climate Change*, eds. Gianfranco Pellegrino, Marcello Di Paola, 1-22. Cham: Springer. https://doi.org/10.1007/978-3-030-16960-2_150-1

Prati, G., Zani, B. (2013) The effect of the Fukushima nuclear accident on risk perception, anti-nuclear behavioral intentions, attitude, trust, environmental beliefs, and values. *Environment & Behavior* 45(6): 782 –798. https://doi.org/10.1177/0013916512444286.

Prelec, D., Wernerfelt, B., Zettelmeyer, F. (1997) The Role of Inference in Context Effects: Inferring What You Want from What Is Available. *Journal of Consumer Research* 24(1): 118–26. https://www.jstor.org/stable/10.1086/209498

R Core Team (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.Rproject.org/.

Reber, A. S. (1993) *Implicit Learning and Tacit Knowledge*. Oxford, UK: Oxford University Press.

Rebonato, R. (2014) A Critical Assessment of Libertarian Paternalism. *Journal of Consumer Policy* 37 (3): 357–96. https://doi.org/10.1007/s10603-014-9265-1.

Rowlands, Ian H., Paul Parker, and Daniel Scott. 2004. Consumer behaviour in restructured electricity markets. *Journal of Consumer Behaviour* 3(3): 273–283. https://doi.org/10.1002/cb.140

Rozin, P., Scott, S., Dingley, M., Urbanek, J.H., Jiang, H., Kaltenbach, M. (2011) Nudge to nobesity I. Minor changes in accessibility decrease food intake. *Judgment and Decision Making* 6(4): 323-332. https://doi.org/10.1017/S1930297500001935

Ryle, G. (1949) *The Concept of Mind*. New York: Barnes and Noble.

Saghai, Y. (2013) Salvaging the concept of nudge. *Journal of Medical Ethics* 39: 487–493. https://doi.org/10.1136/medethics-2012-100727

Schachter, S. (1971) Some extraordinary facts about obese humans and rats. *American Psychologist* 26(2): 129–144. https://doi.org/10.1037/h0030817

Schmidt, A. T. (2019) Getting Real on Rationality—Behavioral Science, Nudging, and Public Policy. *Ethics* 129(4), 511–543. https://doi.org/10.1086/702970

Schmidt, A. T., Engelen, B. (2020) The ethics of nudging: An overview. *Philosophy Compass* 15(4): 1-13. https://doi.org/10.1111/phc3.12658

Schubert, C. (2017) Green nudges: Do they work? Are they ethical? *Ecological Economics* 132: 329–342. https://doi.org/10.1016/j.ecolecon.2016.11.009

Schultz, P.W. (1999) Changing behavior with normative feedback interventions: a field experiment on curbside recycling. *Basic and Applied Social Psychology* 21(1): 25–36. https://doi.org/10.1207/s15324834basp2101_3

Schultz, P. W., Oskamp, S. (1996). Effort as a moderator of the attitude-behavior relationship: General environmental concern and recycling. *Social Psychology Quarterly* 59(4): 375–383. https://doi.org/10.2307/2787078

Schultz, P. W., Zelezny, L. C. (1998) Values and proenviromental behavior: A five-country survey. *Journal of Cross-Cultural Psychology* 29(4): 540‑558. https://doi.org/10.1177/0022022198294003

Schwitzgebel, E. (2024) Introspection. In E. N. Zalta, U. Nodelman (eds.) *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition).
URL = <https://plato.stanford.edu/archives/fall2024/entries/introspection/>.

Sheeran, P. (2002) Intention-behavior relations: A conceptual and empirical review. *European Review of Social Psychology* 12(1): 1–36. https://doi.org/10.1080/14792772143000003

Simon, H. A. (1955) A Behavioral Model of Rational Choice. *Quarterly Journal of Economics* 69(1): 99–118. https://doi.org/10.2307/1884852

Simon, H. A. (1956) Rational Choice and the Structure of the Environment. *Psychological Review* 63(2): 129-38. https://doi.org/10.1037/h0042769

Simon, H. A. (1957) *Models of Man*, New York: John Wiley.

Singh, P., Daga, S. Yadav, K. (eds.) (2024) *Nudging Green: Behavioral Economics and Environmental Sustainability***.** Cham: Springer Nature.

Sloman, S. A. (1996) The empirical case for two systems of reasoning. *Psychological Bulletin* 119: 3-22.

Skinner, B. F. (1938) *The behavior of organisms: An experimental analysis*. New York: Appleton-Century.

Smith, M. B., Bruner, J. S., White, R. W. (1956) *Opinions and personality*. New York: Wiley.

Smith, J. R., Terry, D. J. (2012). Attitudes and behavior: Revisiting LaPiere's hospitality study. In J. R. Smith & S. A. Haslam (Eds.), *Social psychology: Revisiting the classic studies* (pp. 27–41). Thousand Oaks, CA: Sage Publications.

Smith, N. C., Goldstein, D.G., Johnson, E.J. (2013) Choice Without Awareness: Ethical and Policy Implications of Defaults. *Journal of Public Policy and Marketing* 32(2): 159-172. https://doi.org/10.1509/jppm.10.114

Social and Behavioral Sciences Team: Annual Report (Executive Office of the President; National Science and Technology Council, 2015). Available at: https://sbst.gov/download/2015%20SBST%20Annual%20Report.pdf.

Sousa Lourenço, J., Ciriolo, E., Rafael Almeida, S., Troussard, X. (2016) *Behavioural insights applied to policy: European Report 2016*. https://doi.org/10.2760/903938.

Sparks, P., Shepherd, R. (1992) Self-identity and the theory of planned behavior: Assessing the role of identification with the "green consumer". *Social Psychology Quarterly* 55(4): 388-99. https://doi.org/10.2307/2786955

Stanovich, K. E. (1999) *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stanovich, K. E. (2004) *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago, IL: Chicago University Press.

Strack, F., Martin, L. L., Stepper, S. (1988) Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology* 54(5): 768–777. https://doi.org/10.1037/0022-3514.54.5.768

Steele, C. M. (1988) The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in Experimental Social Psychology* 21: 261–302. https://doi.org/10.1016/S0065-2601(08)60229-4

Stern, P.C., Dietz, T., Kalof, L. (1993) Value orientations, gender, and environmental concern. *Environment and Behavior* 25(5): 322-48. https://doi.org/10.1177/0013916593255002

Stern, P. C., Dietz, T., Guagnano, G. A. (1995) The New Ecological Paradigm in social-psychological context. *Environment and Behavior* 27(6): 723–743. https://doi.org/10.1177/0013916595276001

Stern, P. C., Dietz, T., Abel, T., Guagnano, G. A., Kalof, L. (1998) A value-belief-norm theory of support for social movements: The case of environmentalism. *Human Ecology Review* 6: 81–97.

Stone, J., Cooper, J. (2001) A self-standards model of cognitive dissonance. *Journal of Experimental Social Psychology* 37(3): 228–243. https://doi.org/10.1006/jesp.2000.1446

Strick, M., Holland, R. W., van Baaren, R. B., & van Knippenberg, A. (2012) Those who laugh are defenseless: How humor breaks resistance to influence. *Journal of Experimental Psychology: Applied* 18(2): 213–223. https://doi.org/10.1037/a0028534

Sunstein, C.R. (2014) *Why Nudge? The Politics of Libertarian Paternalism*. New Haven: Yale University Press.

Sunstein, C. R. (2016a) *The Ethics of Influence: Government in the Age of Behavioral Science*. New York: Cambridge University Press.

Sunstein, C. R. (2016b) People prefer system 2 nudges (kind of). *Duke Law Journal* 66(1): 121–168.

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

Sunstein, C. R. (2019) *Conformity. The Power of Social Influences*. New York: New York University Press.

Sunstein, C. R., Reisch, L. A. (2013) Green by Default. *Kyklos* 66(3): 398-402. https://doi.org/10.1111/kykl.12028

Sunstein, C. R., Reisch., L. A. (2014) Automatically Green: Behavioral Economics and environmental protection. *Harvard Environmental Law Review* 38: 127-158.

Taleb, N.N. (2010) *The Black Swan. The Impact of the Highly Improbable*. Second Edition. Random House Publishing Group.

Taylor, S. E. (1975) On inferring one's attitudes from one's behavior: Some delimiting conditions. *Journal of Personality and Social Psychology* 31: 126–131. https://doi.org/10.1037/h0076246

Thaler, R. H., Sunstein, C. R. (2003) Libertarian Paternalism. *The American Economic Review* 93: 175-9. https://doi.org/10.1257/000282803321947001

Thaler, R. H., Sunstein, C. R. (2008) *Nudge: Improving Decisions about Health, Wealth and Happiness*. New Haven: Yale University Press.

Thaler, R. H., Sunstein, C. R. (2021) *Nudge. The Final Edition*. New York: Penguin Books.

Todd, P. M., Gigerenzer, G. (2012) Ecological rationality: The normative study of heuristics. In *Ecological rationality: Intelligence in the World*, ed. Peter M. Todd, Gerd Gigerenzer, 487-497. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195315448.003.0142

Valins, S. (1966) Cognitive effects of false heart-rate feedback. *Journal of Personality and Social Psychology* 4: 400–408. https://doi.org/10.1037/h0023791

Vallier, K. (2016) On the inevitability of nudging. *Georgetown Journal of Law & Public Policy* 14: 817–828.

Viale, R. (2022) *Nudging*. Cambridge, MA: The MIT Press.

Vining, J., Ebreo, A. (1992). Predicting behavior from global and specific environmental attitudes and changes in recycling opportunities. *Journal of Applied Social Psychology* 22(20): 1580–1607. https://doi.org/10.1111/j.1559-1816.1992.tb01758.x

von Neumann, J., Morgenstern, O. (1944) *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Vugts, Anastasia, Mariëtte van den Hoven, Emely de Vet, and Marcel Verweij. 2020. How autonomy is understood in discussions on the ethics of nudging. *Behavioural Public Policy* 4(1): 108–123. https://doi.org/10.1017/bpp.2018.5

Wakslak, C. J. (2012) The experience of cognitive dissonance in important and trivial domains: A Construal-Level Theory approach. *Journal of Experimental Social Psychology* 48 (6): 1361–1364. https://doi.org/10.1016/j.jesp.2012.05.011

Wan, C-S., Chiou, W-B. (2010) Inducing attitude change toward online gaming among adolescent players based on dissonance theory: The role of threats and justification of effort. *Computers and Education* 54: 162–168. https://doi.org/10.1016/j.compedu.2009.07.016

Wason, P.C, Evans, St. B. T. (1975) Dual processes in reasoning? *Cognition* 3: 141-54. https://doi.org/10.1016/0010-0277(74)90017-1

Webb, T. L., Sheeran, P. (2006) Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin* 132(2): 249-268. https://doi.org/10.1037/0033-2909.132.2.249

Weigel, R. H., Newman, L. S. (1976) Increasing attitude-behavior correspondence by broadening the scope of the behavioral measure *Journal of Personality and Social Psychology* 33(6): 793-802. https://doi.org/10.1037/0022-3514.33.6.793

Werner, P. D. (1978) Personality and attitude-activism correspondence. *Journal of Personality and Social Psychology* 36(12): 1375–1390. https://doi.org/10.1037/0022-3514.36.12.1375

Wernerfelt, B. (1995) A Rational Reconstruction of the Compromise Effect: Using market Data to Infer Utilities. *Journal of Consumer Research* 21(4): 627–33. https://doi.org/10.1086/209423

Wheeler, G., "Bounded Rationality", *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2020/entries/bounded-rationality/>.

Whitehead, M., Jones, R., Howell, R., Lilley, R., Pykett, J. (2014) Nudging all over the world - assessing the global impact of the behavioural sciences on public policy. Economic and Social Research Council. https://changingbehaviours.files.wordpress.com/2014/09/nudgedesignfinal-1.pdf.

Wicklund, R. A., Brehm, J. W. (1976) *Perspectives on cognitive dissonance*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wicklund, R. A., Cooper, J. Linder, D. E. (1967) Effects of expected effort on attitude change prior to exposure. *Journal of Experimental Social Psychology, 3*, 416–428. https://doi.org/10.1016/0022-1031(67)90006-6

Wilson, T. D. (2002) *Strangers to Ourselves. Discovering the Adaptive Unconscious.* The Belknap Press of Harvard University Press.

Wilson, T. D., Dunn, D. S., Kraft, D., Lisle, D. J. (1989) Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. *Advances in Experimental Social Psychology* 22: 287–343. https://doi.org/10.1016/S0065-2601(08)60311-1

Wilson, T. D., Schooler, J. W. (1991) Thinking too much: introspection can reduce the quality of preferences and decisions. *Journal of Personal and Social Psychology* 60(2): 181-92. https://doi.org/10.1037/0022-3514.60.2.181

Wilson, T. D., Dunn, E.W. (2004) Self-knowledge: its limits, value, and potential for improvement. *Annual Review of Psychology* 55: 493-518. https://doi.org/10.1146/annurev.psych.55.090902.141954

Witte, K., Allen, M. (2000) A meta-analysis of fear appeals: Implications for effective public health campaigns. *Health Education & Behavior* 27(5): 591–615. https://doi.org/10.1177/109019810002700506

World Bank (2015) *World Development Report 2015: Mind, Society and Behavior*. Washington DC: World Bank.

Wright, P. (2002) Marketplace Metacognition and Social Intelligence. *Journal of Consumer Research* 28 (4): 677–82. https://doi.org/10.1086/338210

Wynes, S., Nicholas, K. A., Zhao, J., & Donner, S. D. (2018) Measuring what works: Quantifying greenhouse gas emission reductions of behavioural interventions to reduce driving, meat consumption, and household energy use. *Environmental Research Letters* 13: 1-20. https://doi.org/10.1088/1748-9326/aae5d7