The ethics of digital nudging

PhD Thesis

Paul Christiaan Kuyer, MA, MSc

Dublin City University, School of Theology, Philosophy, and Music

Supervised by Prof Bert Gordijn (Dublin City University), Dr Yashar Saghai (University of Twente), and Daragh O'Brien (Castlebridge)

June 2026

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original and have conformed to the regulations on the use and declaration of Generative AI, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

I hereby certify that no Generative Artificial Intelligence (Gen AI) tools have been used in the creation of the thesis.

Signed: Paul Kuyer ID No.: 19214683 Date: 21-06-2026

Publications

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

Kuyer, P., Shukla, S., O'Brolcháin, F., Doncel, V. R., & Gordijn, B. (2025). The ethics of digital nudging for sustainable energy consumption–A closer look at the EU Green Deal. Technology in Society, 83, Article 102990. https://doi.org/10.1016/j.techsoc.2025.102990

Table of contents

List of figures 7
Abstract 9

Chapter 1 - Introduction
1.1 Research objectives 11
1.2 Methodology 13
1.3 Expected contribution 20
1.4 Structure 23
1.5 Bibliography 29

Chapter 2 – A systematic literature review into the definition and ethics of nudging
2.1 Introduction 32
2.2 Methodology 33
2.3 General results 36
2.4 Ethical issue: Autonomy 48
2.5 Ethical issue: Welfare 57
2.6 Ethical issue: Long-term adverse effects 70
2.7 Ethical issue: Democracy and deliberation 73
2.8 Ethical theories 79
2.9 Discussion 82
2.10 Conclusion 92
2.11 Bibliography 94

Chapter 3 - Digital nudging and the digital nudging imperative of big tech
3.1 Introduction 108
3.2 Definition of digital nudging 109
3.3 Special features of digital nudges 110
3.4 The digital nudging imperative 118
3.5 A user-centric approach focused on secondary education 128
3.6 Conclusion 132
3.7 Bibliography 134

Chapter 4 - A process view on preference shaping for the ethics of digital nudging
4.1 Introduction 140
4.2 Neoclassical welfare economics 144
4.3 Behavioural welfare economics 150
4.4 Behavioural welfare without purified preferences 158
4.5 Conclusion 171

4.6 Bibliography 172

Chapter 5 – How digital nudges distort information
5.1 Introduction 179
5.2 Understanding information's function in decision-making 179
5.3 How digital nudges undermine information 182
5.4 Digital nudges proficient preference shaping 198
5.5 Recommendations 201
5.6 Bibliography 206

Chapter 6 – How digital nudges skew cognitive abilities
6.1 Introduction 211
6.2 Defining cognitive abilities based on dual process theory 211
6.3 System 1 is aligned with lower preferences 221
6.4 Acting for the wrong reasons – cognitive biases 223
6.5 Long-term undermining of system 2 abilities 235
6.6 Recommendations 237
6.7 Bibliography 241

Chapter 7 – How digital nudges undermine willpower
7.1 Introduction 247
7.2 Self-control in the behavioural economic literature 251
7.3 Critiquing the prioritization of system 2 in willpower discussions 257
7.4 Self-control based on resolve 261
7.5 How digital nudges undermine self-control strategies 266
7.6 Recommendations 276
7.7 Bibliography 282

Chapter 8 - Conclusion
8.1 Summary and research objectives 286
8.2 Limitations and suggestions for future research 294
8.3 Bibliography 298

List of figures

Figure 1 PRISMA Table results database search.

Figure 2 Entries ranked on year of publication.

Figure 3 Records per research field.

Figure 4 Nudgers mentioned.

Figure 5 Records categorized on their definition of nudge.

Figure 6 Overview of how records stand towards the proposition that nudges violate autonomy.

Figure 7 Overview of how records stand towards the proposition that nudges improve welfare.

Figure 8 Overview of how records stand towards the proposition that nudges cause long-term adverse effects.

Figure 9 Overview of how records stand towards the proposition that nudges cause issues with democracy and deliberation.

Abstract - The ethics of digital nudging by Paul Kuyer

This dissertation examines the ethics of digital nudging—psychological techniques that predictably alter user behaviour by targeting their shallow cognitive processes. Unlike analogue technologies, digital technologies enable their creators to profit directly from maximizing usage. This creates a digital nudging imperative: an incentive for big tech companies to develop sophisticated influence mechanisms through algorithms and design. This economic imperative which is rooted in the monetization of user attention is the foundation of digital nudges.

The research follows a threefold structure. First, it establishes a technical definition of digital nudging divorced from normative assumptions, positioning these behavioural techniques as intrinsic components of the digital economy's business model. Through a comprehensive literature review, welfare emerges as the most critical ethical concern, as digital users struggle to form and satisfy preferences in environments systematically designed to capture their attention.

Second, this dissertation develops the notion of preference shaping in the context of digital nudges, contrasting it to preference purification, used in conventional behavioural welfare economics. This dissertation identifies two components of proficient preference shaping: experiments in living and reflection. These processes are fundamentally constrained by three limitations that digital nudges deliberately exploit: imperfect information, bounded cognitive abilities, and limited willpower.

Third, the analysis demonstrates how certain digital nudges can systematically undermine welfare by exacerbating these three fundamental limitations—creating information overload and bias, promoting overreliance on shallow cognitive processing, and undermining willpower through stimulating instant gratification. In response, the dissertation recommends developing a user-centric educational approach focused on equipping underage users with skills to navigate digital environments effectively despite these pervasive influences.

By reconceptualizing digital nudges' impact through process rationality rather than outcome rationality, this work contributes to both theoretical discussions in behavioural welfare economics and practical efforts to mitigate the harmful effects of an economic system that profits from compromising user welfare.

Chapter 1 - Introduction

1.1 Research objectives

Digital technologies differ fundamentally from analogue ones because their creators profit directly from capturing and maximizing the usage of their creations. As a result of this incentive structure, tech companies systematically collect behavioural data and refine their algorithms and user interfaces to build increasingly sophisticated psychological influence mechanisms that shape users' digital environments and ultimately, their behaviour. Nudging serves as a particularly valuable tool for these companies, offering subtle yet powerful methods to influence choices. This dissertation investigates the ethics of digital nudges for profit.

Nudging was introduced to the general audience in the seminal book *Nudge* written by behavioural economist Richard Thaler (who later won the Nobel Prize in economics) and legal scholar Cass Sunstein (2008). The book coined the concept nudge and thereby marks the beginning of a rich field of literature on the effectiveness and ethics of nudges but can also be seen as the culmination of several decades of research into dual process theory[1] and behavioural economics[2]. The nudging debate has been shaped by how Thaler and Sunstein originally presented nudges, encompassing two premises.

First, they present nudging as a psychological technique that subtly steers behaviour without altering economic incentives or restricting available choices. It functions by strategically changing how information is presented, targeting cognitive biases. Second, they argue that nudges should be used to steer people

[1] Dual process theories of thinking distinguish between fast, automatic thinking (system 1), and slow and deliberative thinking (system 2) (Evans & Stanovich, 2013). System 1 functions by heuristics and instinct, system 2 by reasoning. Its origins can be traced back to the 1970s (Evans & Stanovich, 2013). Nobel prize winner Daniel Kahneman and Amos Tversky are the most well-known proponents of the theory.

[2] Behavioural economics is a subdiscipline of economics trying to incorporate psychological insights, particularly from dual process theory, into economic science (Anger & Loewenstein, 2012). It focuseson economic behaviour that is not fully rational as commonly understood in economics, but rather on how "people actually behave" (Hargreaves Heap, 2013, p 985) because of the biases and heuristics from system 1.

toward better choices “as judged by themselves” (Thaler & Sunstein, 2008, p. 5), a doctrine its proponents call libertarian paternalism and that many have taken to be an intrinsic element of the definition of nudges (Hansen, 2016; see e.g. Rizzo & Whitman, 2009).

Elaborating on the first premise, nudging is fundamentally a behavioural modification technique grounded in the psychological principles of dual process theory. This theory, developed through decades of research by psychologists including Kahneman and Tversky (cf. Kahneman & Tversky, 1984; Kahneman, 2011), describes human cognition as operating through two systems. A fast, intuitive system (system 1) characterised by Saghai as “shallow cognitive processes” (Saghai, 2013, p. 489); and a slow, deliberative system (system 2). Nudges primarily target system 1, exploiting cognitive biases and heuristics to influence behaviour in predictable ways.

The second premise encapsulates the theoretical commitments and perspectives of the field of behavioural economics. This field posits that agents systematically make welfare-reducing errors due to cognitive biases and limitations in information, cognitive abilities, and willpower, in line with Simon’s (1955) influential work on bounded rationality. It advocates nudging people toward choices they would have made in the absence of these thinking errors to enhance their welfare, a form of ideal preference utilitarianism. This argument gained substantial traction in public policy circles, as governments appeared to be natural candidates for implementing libertarian paternalist nudges. The widespread adoption of this view led to the emergence of behavioural welfare economics, a field dedicated to “reconciling the normative and behavioural economics” (McQuillin & Sugden, 2012, p. 553) (also see Infante et al., 2016; Sugden, 2018; 2020; Desroches, 2020; Dold, 2023 for critical evaluations of the field). The field investigates how to purify the biased preferences of agents to uncover what they would truly prefer under ideal conditions.

To separate normative discussions from conceptual questions, we reject the second premise, i.e. that libertarian paternalism, or any normative commitment, is a necessary element of the definition of nudges. This ensures a cleaner definition based on the techniques used to alter behaviour. We refer to this definition as the technical definition of nudges, and argue it opens the space to investigate private, corporate nudges, which is the scope of this dissertation.

This is important, as the widespread implementation of nudging techniques in digital environments by powerful technology companies (commonly referred to as big tech) raises profound ethical questions that demand rigorous philosophical examination. The ethical implications of digital nudging represent a significant gap in current scholarly discourse. They require careful ethical analysis, particularly when considering the scale and sophistication with which these techniques are deployed. Addressing this important ethical challenge, the main objective of this thesis is to research the ethical implications of digital nudging. To this effect, we advance four subsidiary research objectives.

1) Define the concept of digital nudging
2) Identify the ethical issues associated with digital nudging
3) Analyse the main ethical issues associated with digital nudging
4) Recommend the development of a program to address the main ethical issues associated with digital nudging

1.2 Methodology

1.2.1 Defining digital nudging

Defining digital nudging is a task that should not be underestimated, as nudging is a contentious term with deep conceptual roots in psychology, economic theory and preference utilitarianism that go back to David Hume and Adam Smith (Matson & Dold, 2021; Ashraf et al., 2005). While the application of

nudges in digital environments has not been researched extensively, we can draw upon the broader literature on nudging to develop a robust understanding of digital nudging specifically.

Our investigation begins with a comprehensive literature review examining academic sources published between 2008 and 2020 that address digital nudging, specifically analysing the various definitions employed across the field. Further details on the methodology of this review can be found in chapter 2. Through critical analysis, we identify essential definitional elements and observe that most scholars conceptualize nudges exclusively as benevolent interventions aimed at improving welfare according to purified preferences. We challenge this approach by proposing that nudges should be defined by their psychological mechanisms rather than their intended goals, establishing a technical conception that aligns with dual process theory foundations. Building upon this analysis, we adopt Saghai's (2013) definition as the most robust starting point, with minor modifications for our specific research context.

With this technical definition established, we then specifically apply it to the digital domain, enabling precise identification and analysis of how nudging manifests in digital environments and digital interfaces. We investigate how digital nudging relates to terms like morphing, persuasive design, persuasive profiling, behavioural design, persuasive user interfaces, online behavioural architecture, personalization and digital market manipulation. We then investigate what the special features of digital nudges are and in which domains they are most prevalent. In line with Zuboff's analysis of surveillance capitalism (2019), we argue that digital nudges can only be properly understood in terms of what we call the digital nudging imperative. Digital nudges are a function of the economic revenue models of big tech corporations, which are directly incentivized to maximize user engagement. By establishing these conceptual relationships and situating digital nudges within their digital and economic

context, we develop a comprehensive understanding of what digital nudging means, thereby fulfilling our first research objective.

1.2.2 Identifying the ethical issues associated with digital nudging

To achieve the second objective—identifying the ethical problems associated with digital nudging—we conduct a systematic literature review (the methodological details of which can be found in chapter 2) investigating which ethical issues are raised in the broader nudging literature. Through this review, we find four prominent ethical issues: violations of autonomy, issues concerning welfare, long-term adverse effects, and issues concerning democracy and deliberation.

While all these issues merit consideration, we find that the central issue is that of the normative status of preferences and their formation. This concerns both welfare and autonomy. While some authors have portrayed nudging as creating a tension between autonomy and welfare, others describe the relation between them in a more complex way. In the framework developed in chapter 4, we touch on the relationship between welfare and autonomy, by addressing the way in which preferences should be shaped. We use the ideas of Tiberius, Frankfurt, and J.S. Mill to describe proficient preference shaping.

In addition, we aim to understand nudging in its context of behavioural welfare economics. The issue of welfare connects directly to broader questions in behavioural welfare economics concerning what makes agents better off when they display cognitive biases and inconsistent preferences. When discussing behaviour alteration, digital market manipulation and economic harm, this raises the question what digital users prefer. This discussion provides a natural conceptual bridge to our discussion, which critically examines and extends

behavioural welfare economics to develop a more robust discussion of the ethical implications of digital nudging.

Finally, digital nudges emerge directly from the economic incentives of digital platforms, creating a fundamental tension between corporate profit motives on the one hand and user autonomy and welfare on the other. In the digital economy, characterized by monopolistic competition, network effects, and economies of scale, companies face strong competitive pressures to maximize user attention and engagement. The result on digital users has been referred to as "digital market manipulation" (Calo, 2013, p. 995), bringing the focus on user autonomy and welfare when confronted with digital nudges. This aligns with the issues concerning autonomy and welfare we find in the literature review, including the existence of true preferences and exploitation of error and economic harm.

1.2.3 Analysing a selection of the ethical issues associated with digital nudging

A proper analysis of the ethical issues associated with digital nudging presupposes a precise understanding of welfare and autonomy. These concerns ultimately rely on what agents prefer and choose, and on the normative demands that should be placed on the decision-making processes leading to those choices.

We approach these ethical issues from the standpoint of the agent herself, focusing not on abstract accounts of welfare but on how individuals ought to navigate their lives through decision-making over time. In this spirit, we draw on Valerie Tiberius' (2008) proposal of "a first-personal, process-based account of how to live" (p. 3), as opposed to an impersonal, goal-based theory of the good life. To put it another way, we begin with the question 'How should I live?' instead of questions like 'What is a happy life?' or 'What is a good life for a human being?'" (p. 3).

Tiberius asks what the appropriate first-person perspective is and argues there are two obvious answers to this question. First, the appropriate point of view could be the rational, reflective point of view. On this account, a life well lived means agents follow a life in which they act in accordance with plans and ideas they reach after careful and rational deliberation. Tiberius notes that this account is at the centre of a broad philosophical tradition, starting with Plato's metaphor of the chariot and the charioteer in which the emotional, intuitive part of the soul (the chariot) must follow the charioteer, which symbolizes the rational part of the soul.

However, Tiberius notes that "recent investigations in empirical psychology show us that the self-conscious, rational processor is more fallible than we imagined" (2008, p. 4). This is indeed the conclusion that is at the foundation of the literature of behavioural economics and nudging (Kahneman, 2011; Thaler, 2015); Kahneman's research divides human cognition into two selves, one fast and intuitive (system 1), and one slow and deliberative (system 2), demonstrating that our rational faculties are far less in control than we tend to assume.. Tiberius summarizes this body of research by stating that system 2, the deliberative self, uses biased, incomplete, and fallible thinking. In addition, she states, the intuitive system 1 is not easily controlled as system 1 influences a great deal of human unconsciously.

Tiberius suggests that Plato's metaphor of the chariot and the charioteer should perhaps be replaced by Jonathan Haidt's version of the elephant and the rider. Haidt, who investigates the psychology of human morality and intuition, favours the metaphor of the elephant (intuition), which chooses its own direction the majority of the time, which leaves the rider (deliberation) in a position of following most of the time, providing post-hoc rationalizations for behaviour rather than genuinely controlling behaviour (Haidt, 2005; 2012).

If we understand agents to be more irrational, then how should we proceed with the question: how should one live? Tiberius rejects the idea that the

intuitive self must be completely in control, as that self is also fallible, and since humans are reflective beings. She also rejects the approach to define an ideal standard of rational reflection—one that is entirely free from the biases, errors, and constraints that characterize ordinary human thinking—and then to identify a person's good with the life she would choose under those idealized conditions (Railton, 1986; Griffin, 1986). This approach has intuitive appeal when we are evaluating what is good for a person from an external standpoint. However, it is far less convincing as guidance for living from the first-person perspective, she argues. Since human beings are neither perfectly rational nor ever likely to become so, it offers little practical direction to suggest that we should simply choose whatever our ideally reflective selves would choose. Moreover, given the specific limits of our actual deliberative capacities, attempting to model our decision-making on a hypothetical perfectly rational agent may even be counterproductive. Mistakenly assuming that an elephant can be steered may exercise less effective control than realistically understanding we are guiding a powerful and largely autonomous elephant.

Tiberius instead proposes a strategy wherein we answer the question how to live by training the reflective system 2 in a way that it can coexist and function well together with the intuitive system 1 and innate emotions and desires. This strategy therefore begins from an attitude of humility about the capacities of reflective deliberation, combined with a recognition of the practical importance of non-reflective processes in guiding behaviour. Acknowledging the systematic limits and biases of reflection does not entail accepting the outputs of our actual deliberation at face value. Rather, it motivates an approach to improving decision-making that does not rely on comparison with an unrealistically idealized, fully rational agent. Given our knowledge concerning predictable errors in explicit reasoning and heuristic success, reflective improvement must be understood in more pragmatic terms. On this view, reflection can be trained and supported by cultivating stable cognitive dispositions, decision habits, and

self-regulatory practices that work with—rather than against—the constraints of human psychology. Improving reflection, then, is not a matter of approximating ideal rationality, but of enhancing the reliability of real-world decision processes under conditions of bounded rationality.

To base our investigation on psychological research and behavioural accounts, we stress three fundamental limitations to human rationality: bounded information, cognitive abilities, and willpower. Acknowledging these, forms the basis of our account of how to live from a first-person view. System 2 and rational reflection are neither the sovereign controller of action nor an idealized standard against which actual choices are to be measured. Rather, it is a capacity that can be developed, supported, or undermined under conditions of bounded rationality. This perspective shifts normative attention away from detached assessments of what agents would choose under ideal rational conditions and toward the quality of the decision-making processes available to them in practice. The central ethical question thus becomes whether choice architectures support or erode agents' capabilities to reflect, choose, and revise their commitments over time. In chapter 4, we argue for such a process-oriented perspective, evaluating digital nudges by their impact on these capabilities.

1.2.4 Recommending the development of a program to address the selected ethical issues associated with digital nudges

Based on the ethical analysis of each fundamental limitation, we then translate these analyses into theoretical recommendations that suggest directions for future educational research aimed at mitigating the harmful effects of digital nudges on welfare. We first argue for a user-centric approach, focusing the recommendations, like the ethical issue and moral framework, on users. From the three obvious stakeholder groups (users, digital designers, and governments), we argue users received the least attention in the literature, while

users are a promising source of positive change as both big tech corporations and governments face significant constraints in addressing these ethical issues. Digital nudging is vital to the business models of tech companies, making them unlikely to implement meaningful reforms without external pressure. Many governments face jurisdictional challenges, as the largest digital nudgers are primarily based in the United States.

In contrast, users themselves have both the motivation and capacity to mitigate the harmful effects of digital nudges when equipped with appropriate knowledge and skills. Therefore, more attention for the role of users is needed in addition to the valuable work addressing designers and governments. A focus on education is particularly promising because children and teenagers are growing up immersed in digital environments saturated with behavioural influences. Young people represent especially heavy digital users. US teenagers spend approximately seven hours daily with screens, excluding schoolwork (Robb, 2019). In addition, they are at a formative stage where interventions can be most effective, and the educational system offers a means of reaching them. By providing these young users with the necessary cognitive tools, we can enhance their digital resilience and strengthen their capacity to shape preferences proficiently despite the influence of digital nudges.

In chapter 3, we address several counterarguments against a user centric approach, namely that it unduly and unfairly burdens users, and that (some) users are incapable of learning strategies against digital nudges.

Our recommendations address the three fundamental limitations that digital nudges exploit. For each limitation, we propose theoretical educational approaches that suggest directions for future research into practical educational interventions.

1.3 Expected contribution

1.3.1 Academic relevance

This dissertation contributes to the academic literature in three ways. First, it provides a systematic literature review on the ethics of nudging, creating a comprehensive overview of existing scholarship. Our comprehensive analysis identifies and categorizes the main ethical concerns across multiple disciplines, providing clarity on the state of the debate and offering a sound foundation for this dissertation but also for future contributions in the field.

Second, this dissertation improves conceptual clarity for the term digital nudge, which is a novel concept that is used differently across various research fields and disciplines. By distinguishing digital nudges from other behavioural modification techniques while aligning it with its roots in cognitive psychology and dual process theory specifically, we establish a precise definition that can guide future research.

Third, this dissertation proposes a new avenue for behavioural welfare economics based not on outcome rationality and preference purification but on process rationality and proficient preference shaping. This is a major development in the emerging field of behavioural welfare economics as it clarifies how individuals try to judge and pursue their own welfare and which obstacles, in terms of digital nudging, they may face

By critiquing the preference purification approach and instead focusing on the processes of preference shaping and satisfaction, we offer a more realistic framework that acknowledges the inevitable limitations humans face. Our application of this framework to digital nudges demonstrates its analytical power and practical relevance, particularly in evaluating how digital nudges impact an agent's ability to navigate the three fundamental limitations of information, cognitive abilities, and willpower. In addition, we address the concept of preference shaping, which is a crucial element when rejecting the

assumption of complete and consistent preferences. This discussion offers a meaningful starting point for future research and contributes to the advancement of behavioural welfare economics as an emerging field.

1.3.2 Social relevance

Research into the ethics of digital nudging is socially relevant as behavioural influence techniques are widely adopted across both public and private sectors. Governments worldwide have established specialized behavioural insights teams to implement nudge practices in public policy (cf. Halpern, 2015), raising questions about citizen autonomy and political legitimacy. In the private sector, corporations leverage behavioural research combined with large amounts of user data to steer consumer behaviour in digital environments, prompting concerns about manipulation and welfare impacts.

Specifically, this dissertation contributes socially in four ways. First, it provides an understanding of what digital nudging entails and how it is driven by what we term the digital nudging imperative, a necessary result of the big tech revenue model. This allows for a better understanding of digital nudges and where they come from. By explaining the competitive dynamics of the digital economy and its winner-take-all nature, we demonstrate the digital nudging imperative; why digital nudges are inevitable within the current business model of the attention economy. This insight can aid important social debates on the role of big tech in society and concerning the development of digital culture and digital practices.

Second, we identify specific welfare harms caused by digital nudges through their exacerbation of the three fundamental limitations. By showing how digital nudges create information overload and bias, promote overreliance on system 1 thinking, and undermine self-control strategies, we establish a clear connection between digital nudging and compromised welfare. This analysis provides policy makers, educators, and citizens with a framework to recognize harmful digital design patterns and understand their effects, enabling more

informed decisions about digital tool usage and supporting the development of personal strategies to mitigate these harms.

Third, by providing a novel direction for behavioural welfare economics based on an understanding of the fundamental limitations, we offer a framework that can guide policy development. While not the explicit focus of this dissertation, our analysis uses the literature on nudging to show that governmental policies aimed at improving the welfare of citizens can be improved by adopting a focus on process rationality and proficient preference shaping. These insights can inspire policy makers to focus on creating beneficial environments and conditions that aid information, cognitive abilities, and willpower, rather than determining what citizens prefer.

Fourth, this research provides a framework in terms of the user-centric approach focused on education to mitigate the adverse effects of digital nudging. Specifically, it offers recommendations for educators and educational policy makers on which skills should be taught to help teenagers navigate digital environments, process information critically, strengthen cognitive abilities, and develop effective self-control strategies.

1.4 Structure

This thesis has a threefold structure. First, we define digital nudges, and investigate which ethical issues they raise. We argue for conceptualizing digital nudges as an intricate component of the digital economy and argue that welfare is the most important ethical issue to investigate. Second, we develop a moral framework based on behavioural welfare economics which recognizes agents must both proficiently shape and satisfy their preferences to attain welfare. In doing so, they face three fundamental challenges, namely information, cognitive abilities, and willpower. Third, we apply the moral framework to investigate how digital nudges impact welfare through their impact on the three fundamental

limitations and formulate recommendations on how agents can mitigate the harmful effects of digital nudges.

Chapter two presents the systematic literature review on the ethics of nudging. We found 131 relevant academic sources and evaluated three elements for each: 1) which definition of nudging does the source use, 2) which ethical issues does it identify and 3) which moral theory does it use to assess the ethical issues. With respect to 1) we find that most sources use Thaler & Sunstein's libertarian paternalist definition of nudges but that an important minority instead subscribes to Saghai's (2013) technical definition. With respect to 2) we also find that there are four prominent ethical issues in the literature: violations of autonomy, issues concerning welfare, long-term adverse effects, and issues concerning democracy and deliberation. With respect to the third element 3) we find that most of the literature describes the ethics of nudging in terms of libertarian paternalist nudges, which have the goal of improving the welfare of the nudgee. From this perspective, the ethics of nudging is usually presented as a tension between improving welfare and respecting autonomy. Welfare is frequently understood in terms of Thaler and Sunstein's proposal of ideal preference utilitarianism. The chapter concludes with a discussion in which we argue that nudges are best conceptualised in technical terms, devoid of normative commitments. Doing so provides a space to investigate private digital nudges. Finally, we argue that welfare is the most important issue to investigate.

Chapter three transitions from nudges to their specific application in digital environments. We distinguish digital nudges from related behavioural modification terms such as "online manipulation" (Susser et al., 2019, p. 2), "persuasive (digital) technology" (Fogg, 2003, p. 1), and "persuasive digital advertising" (Williams, 2017, p. 129; Aylsworth, 2020, p. 3) before identifying five defining features: they operate through digital interfaces (making them

visually appealing, interactive, and unconstrained by time and space); they are measurable (allowing precise tracking of users); they are adaptable (enabling cheap, frequent changes); they are personalized (increasing effectiveness); and they remain largely invisible to users (as underlying algorithms are difficult to observe).

We then argue that digital nudges can only be understood as the result of what we call the digital nudging imperative. The dynamics of the digital economy, in which a few dominant tech corporations compete for the time and attention of users, necessitate the heavy use of digital nudges. Due to monopolistic competition, network effects and economies of scale and scope, the digital economy is characterized by a winner take all dynamic and the business model of these companies relies on capturing attention. As a result, digital nudges are a necessity for these firms to compete. This becomes clear mainly in the domains of operating devices (such as Android and MacOS), online advertising (including social media), and e-commerce.

We argue that combating harmful digital nudges requires a user-centric approach, emphasizing education that empowers users to protect themselves. Developing educational programs for teenagers is crucial, as they are both frequent users of digital environments and vulnerable to manipulation, yet capable of learning protective skills.

Chapter 4 discusses the notion of preference shaping, using Mill, Tiberius, and Frankfurt. We analyse and critique the concept of welfare used in behavioural welfare economics. To understand this emerging field, it is essential to first grasp its roots in neoclassical welfare economics. We then examine the current state of behavioural welfare economics, which predominantly characterizes agents as frequently making welfare-reducing errors due to thinking biases and three fundamental limitations: information, cognitive abilities, and willpower.

While we agree that agents face these three fundamental limitations, preventing them from ever attaining complete and consistent preferences, we argue that preference purification is problematic and fruitless as a direction for the field of behavioural welfare economics. Instead, recognizing these inevitable limitations should lead us to acknowledge that incomplete preferences necessitate a focus on attainable processes. These processes form the core of our framework: proficient preference shaping. Agents must first shape their preferences before meaningfully satisfying them, and proficiency in both processes is essential for welfare.

Drawing from John Stuart Mill and Valerie Tiberius, we identify two critical components of proficient preference shaping: conducting experiments in living (testing different lifestyles through direct experience) and reflection (developing meta-preferences that guide preference shaping). These components enable agents to navigate the inevitable inconsistencies and incompleteness in their preferences, transforming these incomplete preferences into opportunities for learning and development.

The three fundamental limitations directly impact preference shaping and satisfaction processes, thereby affecting welfare. Our moral framework addresses the ethics of digital nudging by examining how digital nudges exacerbate these limitations and affect preference shaping. The subsequent three chapters address each limitation in turn.

Chapter 5 investigates how digital nudges impact the fundamental limitation of information. Agents can never have or process all the information relevant to their decisions. Digital nudges aggravate this limitation as they algorithmically manipulate information environments, causing information overload. In addition, digital nudges bias information through filtering and personalization. Due to the economic imperative of the digital economy, digital platforms favour information that captures and maintains the attention of users.

In line with Hills (2019), we argue that as a result, the information on these platforms is systematically biased towards belief-consistent, social, negative, predictive, and empty information. As a result, agents have greater difficulty understanding the alternatives available to them. In addition, the biased information environment distorts feedback mechanisms and thereby proficient preference shaping.

To mitigate the harmful effects of digital nudges on information, we propose the following four recommendations, targeting teenagers through an educational program developed in line with the proposed user-centric approach:

- Develop curricula on understanding digital information environment dynamics
- Teach mindful information consumption techniques to prevent overload
- Guide students in seeking diverse and deep information sources
- Train students to customize digital tools and manage notifications

Chapter 6 examines how digital nudges impact the fundamental limitation in cognitive abilities. Building on our definition of nudges as interventions targeting shallow cognitive processes, we analyse these processes through the lens of dual process theory—distinguishing between automatic system 1 and deliberative system 2 thinking. We argue, in line with Stanovich (2004; 2011), that system 1 is more aligned with evolutionary needs and lower preferences. While both systems serve important functions, digital nudges promote excessive reliance on system 1, causing agents to act based on biases related to myopia, cognitive fluency, risk aversion, and conformity. Furthermore, this overreliance potentially gradually erodes system 2's reflective capacities and undermes the reflective abilities essential for proficient preference shaping. To mitigate these effects, we offer the following educational recommendations:

- Integrate theoretical knowledge about digital nudges and cognitive abilities into curricula

- Develop programs focused on strengthening cognitive abilities and reflection
- Demonstrate techniques for introducing friction into digital interactions
- Practice critical thinking tools in digital contexts

Chapter 7 examines how digital nudges impact the fundamental limitation of willpower. Building on our analysis of willpower as the widespread violation of temporal consistency, we explore how digital nudges exploit the internal conflict between an agent's two conceptual selves—the short-term self, associated with system 1 thinking and short-term preferences, and the long-term self, associated with system 2 thinking and long-term preferences. While the dominant view in behavioural economics conceptualizes willpower as the suppression of system 1 impulses by system 2, we argue for a more balanced perspective that acknowledges both systems' contributions to self-control. Drawing from Ainslie's (2021) work, we identify three key self-control strategies: situational control, suppression, and resolve. Digital nudges undermine these strategies by limiting environment control, increasing the need for self-control through ubiquitous temptations, and interfering with the intertemporal bargaining process needed for resolve. This interference prevents agents from achieving alignment between their short-term and long-term preferences, thereby compromising proficient preference shaping and welfare. To mitigate these effects, we suggest the following educational recommendations:

- Teach students to recognize the internal conflict between their two selves
- Train students in applying situational control strategies
- Explain the limitations of suppression as a self-control method
- Teach students to resolve temptations through pattern recognition and habit formation

Chapter 8 summarizes the main findings of this investigation, discusses its limitations and suggests avenues for future research.

1.5 Bibliography

Ainslie, G. (2021). Willpower with and without effort. *Behavioral and Brain Sciences*, *44*, e30. https://doi.org/10.1017/S0140525X20000357

Angner, E., & Loewenstein, G. F. (2012). Behavioral economics. In U. Mäki (Ed.), *Handbook of the philosophy of science: Philosophy of economics* (641-690). Elsevier.

Ashraf, N., Camerer, C. F., & Loewenstein, G. (2005). Adam Smith, behavioral economist. *Journal of Economic Perspectives*, *19*(3), 131–145.

Aylsworth, T. (2020). Autonomy and manipulation: Refining the argument against persuasive advertising. *Journal of Business Ethics*, 1–11.

Calo, R. (2014). Digital Market Manipulation. *The George Washington Law Review*, *82*, 58.

Dold, M. (2023). Behavioural normative economics: Foundations, approaches and trends. *Fiscal Studies*, 44(2), 137–150. https://doi.org/10.1111/1475-5890.12340

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fogg, B. J. (2003). Persuasive Technology: using computers to change what we think and do. Morgan Kaufmann Publishers Inc.

Haidt, J. (2005). *The happiness hypothesis: Finding modern truth in ancient wisdom*. Basic Books.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Halpern, D. (2015). Inside the nudge unit: How small changes can make a big difference. Random House.

Hansen, P. G. (2016). The Definition of Nudge and Libertarian Paternalism: Does the Hand Fit the Glove? *European Journal of Risk Regulation*, 7(1), 155–174. https://doi.org/10.1017/S1867299X00005468

Hargreaves Heap, S. P. (2013). What is the meaning of behavioural economics? *Cambridge Journal of Economics*, *37*(5), 985–1000. https://doi.org/10.1093/cje/bes090

Hills, T. T. (2019). The Dark Side of Information Proliferation. *Perspectives on Psychological Science*, *14*(3), 323–330. https://doi.org/10.1177/1745691618803647

Infante, G., Lecouteux, G., & Sugden, R. (2016). Preference purification and the inner rational agent: A critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, 23(1), 1–25. https://doi.org/10.1080/1350178X.2015.1070527

Griffin, J. (1986). Well-being: Its meaning, measurement, and moral importance. Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341.

Matson, E. W., & Dold, M. (2021). The Behavioral Welfare Economist in Society: Considerations from David Hume. *Review of Behavioral Economics*, *8*(3–4), 239–258. https://doi.org/10.1561/105.00000141

McQuillin, B., & Sugden, R. (2012). Reconciling normative and behavioural economics: The problems to be solved. *Social Choice and Welfare*, 38(4), 553–567.

Mill, J. S. (2001). *On liberty* (Original work published 1859). Batoche Books.

Nys, T. R., & Engelen, B. (2017). Judging Nudging: Answering the Manipulation Objection. Political Studies, 65(1), 199–214. https://doi.org/10.1177/0032321716629487

Railton, P. (1986). Moral realism. The Philosophical Review, 95(2), 163–207. https://doi.org/10.2307/2185589

Rizzo, M. J., & Whitman, D. G. (2009). Little brother is watching you: New paternalism on the slippery slopes. *Arizona Law Review*, 51(3), 685-740.

Robb, M. (2019). *Tweens, teens, and phones: What our 2019 research reveals*. Common Sense. https://www.commonsensemedia.org/kids-action/articles/tweens-teens-and-phones-what-our-2019-research-reveals

Saghai, Y. (2013). Salvaging the concept of nudge. *Journal of Medical Ethics*, *39*(8), 487–493. https://doi.org/10.1136/medethics-2012-100727

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

Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. University of Chicago Press.

Stanovich, K. E. (2011). *Rationality and the reflective mind*. Oxford University Press.

Sugden, R. (2018). The community of advantage: A behavioural economist's defence of the market. Oxford University Press.

Sugden, R. (2020). Normative economics without preferences. *International Review of Economics*, 1–15.

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Online Manipulation: Hidden Influences in a Digital World. *Georgetown Law Technology Review*, *4*, 1.

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Penguin.

Thaler, R. H. (2015). Misbehaving: The making of behavioural economics. Penguin Books.

Tiberius, V. (2010). The reflective life: Living wisely with our limits. OUP Oxford.

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

Williams, J. (2017). *Freedom and persuasion in the attention economy* (Doctoral dissertation). University of Oxford.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

Chapter 2 – A systematic literature review into the definition and ethics of nudging

2.1 Introduction

In this chapter, we systematically review the academic literature on the ethics of nudging, as well as the definition of nudging in depth. Both nudging and its libertarian paternalist justification are controversial and have sparked a vast literature on the ethics of nudging that encompasses arguments from economics, law, politics, public policy, and philosophy. Against this backdrop, the research objectives of this study are threefold: 1) to examine and clarify the definitions of nudging employed in the scholarly discourse; 2) to identify and analyse the ethical issues that have been recognized in relation to nudging practices; and 3) to investigate the ethical theories that have been applied in the debate surrounding nudging. Hereby, we partially address the first two objectives of this thesis i.e. 1) define the concept of digital nudging; and 2) identify the ethical issues associated with digital nudging. In chapter 3 we expand on these findings by adding an explicit focus on digital nudges.

There have been several reviews on the ethics of nudging literature. Schmidt and Engelen (2020) present an overview of the arguments in the debate on the ethics of nudging, which is valuable and of high quality. However, they do not systematically review the literature. They simply focus on concisely presenting the argument in favour of and against nudging. Second, Vugts et al. (2020) do systematically review the literature, albeit only with respect to the ethical issue of autonomy, and only on a selection of 33 articles. We use their valuable insights to structure our own discussion of autonomy. Third and last, Congiu and Moscati (2022) review the literature on three points: the definition of nudging, the justifications of nudging, and the effectiveness of nudging. Our review shares some overlap mainly on the point of the definition. Aside from the

definition, their review engages more with the empirical literature, where we focus on the philosophical literature, as our focus is on ethical issues.

This literature review proceeds in three parts. First, we discuss how the articles included in the review were selected and analysed in the methodology section. Subsequently, in the results section we present the answers to the three research questions above. Finally, the discussion critically reflects on these findings.

2.2 Methodology

A preliminary exploration of the literature in three databases[3] on the concept digital nudging reveals that digital nudging is a relatively new concept which has not been researched extensively. It can be concluded that the literature on digital nudging is still a small and emerging field. A systematic literature review with these inclusion criteria would be fruitless.

The debate on the ethics of nudging on the other hand, is both developed and mature. There is a lively and broad research field into the ethics of nudging. The arguments made in debates on the ethics of nudging can be applied to the more specific field of the ethics of digital nudging. Therefore, I undertake a systematic literature review on the ethics of nudging, rather than on the ethics of digital nudging. Given the fact that the literature on the ethics of digital nudging is still small and developing, an analysis of the ethics of nudging is the best possible starting point.

Since we expect to find discussions of the ethics of nudging in various disciplines, this review uses three disciplinary databases: EbscoHost (academic search complete), Web of Knowledge (core collection) and Scopus. Together, these databases contain publications from professional societies and reports, as

[3] Ebsco, Scopus and Web of Knowledge

well as peer-reviewed journals and online resources, covering a wide variety of research disciplines.

Two sets of search terms are used. The first set consists of the search term nudg*" (to include nudge, nudges, nudging, etc.) and the related phrases "choice architect*" and "libertarian paternal*". The search terms in the second set capture the concept of ethical issue, as defined previously. These include "ethic*", "moral*", "valu*", "norm*" and "virtue*". Together, these sets form the following search string: "nudg*" OR "choice architect*" OR "libertarian paternal*") AND ("ethic*" OR "moral*" OR "valu*" OR "norm*" OR "virtue*". Entering these search terms in the selected databases and excluding all languages other than English yields 488 results in Ebsco, 1334 in Web of Knowledge and 1248 in Scopus (see the PRISMA table in Figure 1 below). After filtering out duplicates, the total of unique identified records is 1693.

In the screening stage, we assess the eligibility of all records. The sole inclusion criterion for eligibility is that a record must significantly discuss at least one of the three topics of interest: 1) the definition of nudge; 2) the ethical issues with nudging; or 3) the methods to evaluate nudge. The records can be divided into four categories: a) peer reviewed journal articles; b) books; c) book chapters or sections; and d) commentaries, reviews, editorials, and responses. To assess the eligibility of a peer reviewed article, we read the title, abstract and bibliography. For books, we read the title, table of contents, and the introduction. For book chapters that have an abstract (commonly those in edited collections), we read the title, abstract and bibliography while for those that do not have an abstract, we read the introduction instead. Since most commentaries do not contain substantial arguments, they were excluded. Filtering the records based on these criteria leaves a total of 305 unique results. Subsequently, we read the screened records in full to definitively assess their relevancy. We exclude

another 174 records as they do not meet the defined inclusion criterion, which leaves 131 records.

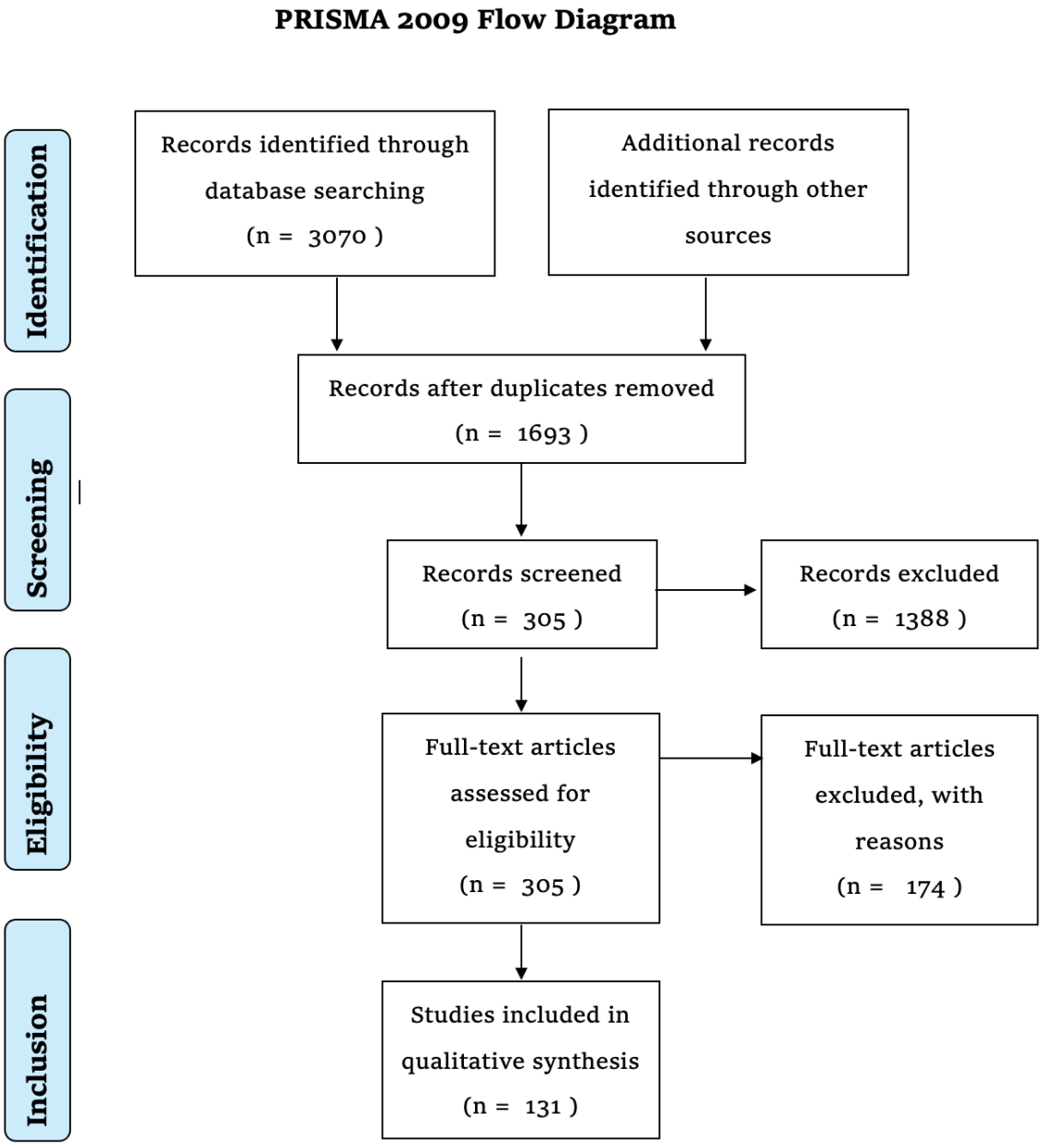


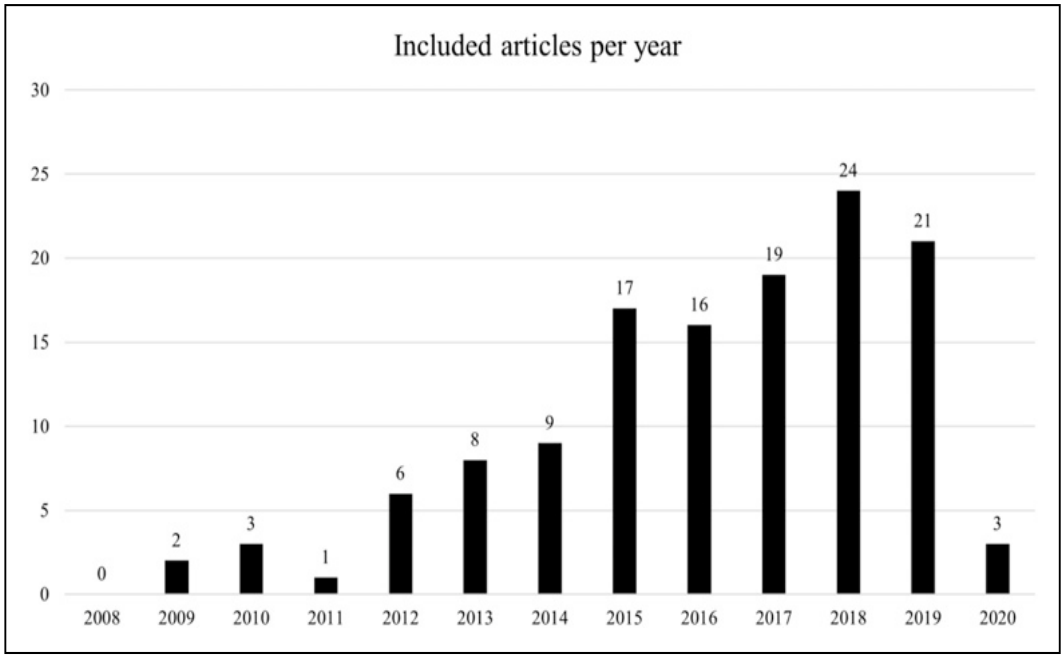
Figure 1 - PRISMA Table Results database search.

We analysed these records by annotating them. All sections that go into either the definition of nudge or the morality of nudges, broadly conceived, were marked. For those marked sections that concern the morality of nudging was then asked: is the author raising an ethical issue here? If the answer was affirmative, it was asked whether this ethical issue was seen previously in the review, or if the ethical issue was new to the review. In this way, we uncovered four ethical issues, as presented in the following section.

2.3 General results

In this result section, we start with a general overview of the literature. We then present our findings with respect to definitions, four ethical issues with, and theories applied to nudging. The literature on the ethics of nudging commenced with the publication of the book *Nudge* in 2008. As becomes evident from Figure 2, the number of contributions has gradually been growing since then (the literature review includes entries until 2020; the thesis until 2024).

Figure 2 - Entries ranked on year of publication.



2.3.1 Research fields

The articles can be categorized into their respective research fields (Figure 3). In the largest research field, healthcare, the morality of nudges aimed at improving health is discussed. This domain includes discussions about using nudges to increase organ donation, stimulate healthy living (e.g., by discouraging drinking and smoking (Quigley, 2014; Ram-Tiktin, 2018; Schmidt, 2016), stimulating healthy nutrition (Bonotti, 2015; Korthals, 2015) and physical exercise), increase vaccination rates (Navin, 2017), paediatric practice (M'hamdi et al., 2017; Navin, 2017; Stanak, 2019) decrease winter deaths (Allmark & Tod, 2014), steer patient choices during important health decisions such as to operate or not (Epstein, 2017; Gorin et al., 2017), and steer people to register for pre-emptive screening of diseases (Damhus et al., 2018; Hofmann & Stanak, 2018). It also includes questions surrounding informed- and presumed consent (Whyte et al., 2012; Cohen, 2013; Gelfand, 2016; Mehlman et al., 2018) and the question whether family members of a patient could be nudged (Blumenthal-Barby & Opel, 2018; Chandler & Gruben, 2016; Sharif & Moorlock, 2018). The second field contains records discussing politics, legislation, and law, focusing on the

morality of nudges by the government or the legislator to improve citizen well-being. In the third field, philosophy, the morality of nudges is discussed in abstracto, without explicitly applying them to any area of empirical research. The fourth research field tackles the morality of nudges in technology. This field mainly focuses on digital technology, but it also includes records on the question whether robots could nudge (Borenstein & Arkin, 2016, 2017; Burr et al., 2018), the relationship between technology as a shaping factor and nudging as a shaping factor (Dorrestijn & Verbeek, 2013) and adaptive or personalised choice architectures (Susser, 2019). In the fifth field, economics, sources discuss the formulation of economic concepts such as preferences, rationality and welfare and their respective normative implications. Next, the environmental studies domain includes all nudges that aim to steer people towards greener choices (see for example Hilton, et al., 2018). The last fields consist of one or two papers: nudging decisions in information security (Renaud & Zimmerman, 2018), nudging in marketing (Smith et al., 2013), nudging as financial regulation (White, 2017), nudging charitable donations (Hobbs, 2017), and nudging safety in the process industry (Lindhout & Reniers, 2017).

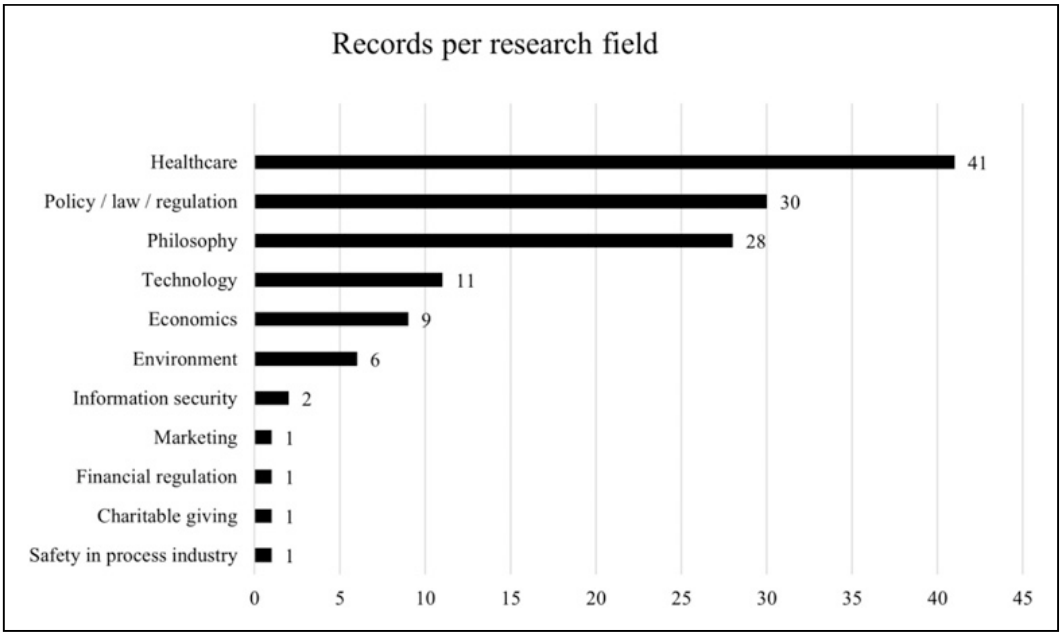


Figure 3 - Records per research field.

2.3.2 Nudger

Of each record we have identified who are considered to be the nudgers or choice architects. As can be seen from Figure 4, the government is by far the most frequently mentioned candidate to perform nudging. Some authors even define nudging as an act performed by the government, or as a policy tool (Grüne-Yanoff, 2018; Häußermann, 2020). This can be attributed to Thaler and Sunstein's (2008) plea for governmental nudging to improve citizen wellbeing. While a large part of the literature focuses on the government as nudger, nudges need not necessarily be performed by the government, as private actors can

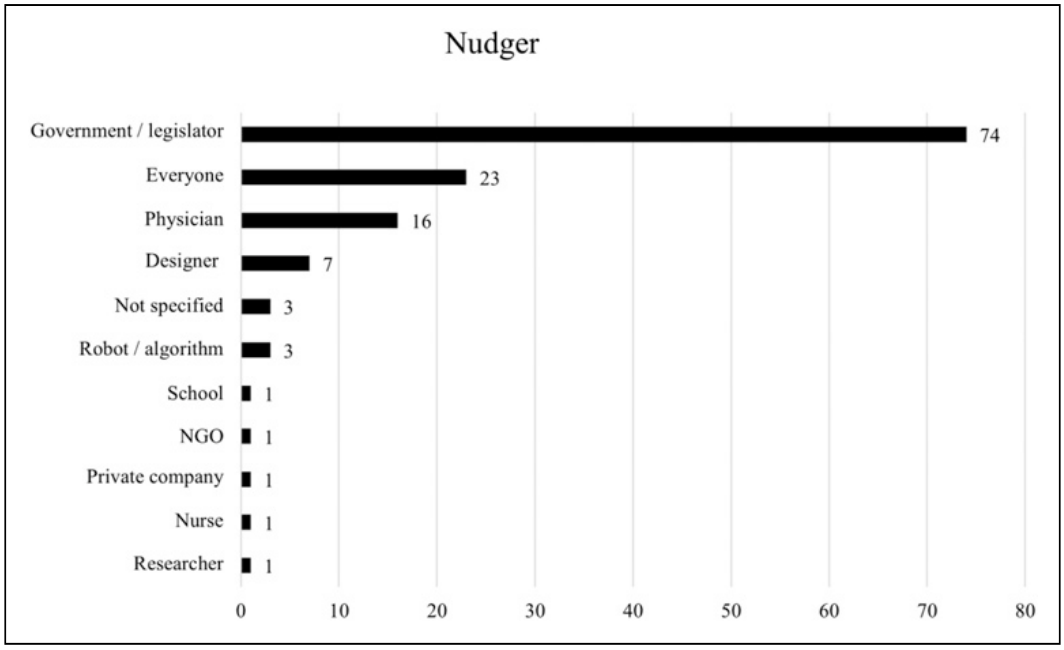


Figure 4 - Nudgers mentioned.

nudge too. Often mentioned candidates other than the government are physicians, technological designers, and engineers.

2.3.3 Definitions of nudging

We first present Thaler and Sunstein's (2008) definitions of nudge and the related concept libertarian paternalism. We then cover several critiques of these definitions. Finally, we will review alternatives to Thaler and Sunstein's definitions of both nudge and libertarian paternalism.

Thaler and Sunstein's definition of nudge

The term 'nudge' was coined by Thaler and Sunstein (2008) who define it as: "any aspect of the choice architecture that alters people's behaviour 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" (Thaler & Sunstein, 2008, p. 6).

This definition is considered the original definition and is used frequently in the literature. Of all records, 37% either subscribe to this definition or

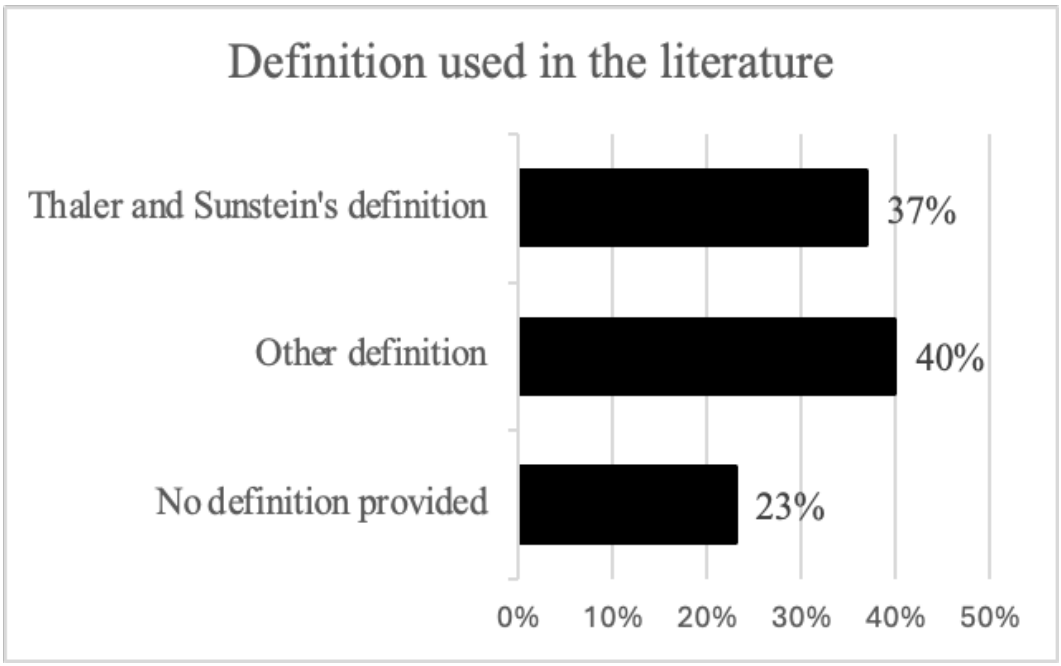


Figure 5 - Records categorized on their definition of nudge.

rephrase it without making significant alterations. Hence, this definition can be considered the standard. 23% of the records do not provide a definition; sometimes these sources focus on libertarian paternalism instead. The remaining 40% provide a different definition, as will become clear in the subsequent section.

Thaler and Sunstein's definition of libertarian paternalism

Libertarian paternalism is the idea that since the choices of people are frequently irrational, they should be steered toward better outcomes without restricting their freedom of choice (Sunstein & Thaler, 2003). The idea of helping people to make better choices makes it a paternalist theory. Thaler and Sunstein argue outcomes are better when they make the agents who are steered "better off, as judged by themselves" (Thaler & Sunstein, 2008, p. 5). It is libertarian since it does not ban or tax bad choices, but it nudges people in the right direction. People are free to resist nudges if they want, leaving the ultimate choice to the choosers. It must be added that for Thaler and Sunstein (2008) governments are not the only actors that can act in a libertarian paternalistic

way. Other institutions such as NGOs and corporations can also act in a libertarian paternalistic manner.

Critique of Thaler and Sunstein's definition of nudge

It is argued that Thaler and Sunstein's definition is unclear (Floridi, 2016) problematic and vague (Baldwin, 2014) and incomplete (Hansen, 2016). Some authors complain Thaler and Sunstein define nudge by example, while some of the examples they list are inconsistent with their own definitions (Selinger & Whyte, 2011; Hausman & Welch, 2010). The ambiguity over the concept nudge is aggravated due to its relatedness to the concept libertarian paternalism. Some authors focus on the definition of 'nudge', while others take 'libertarian paternalism' as their target concept. To make matters worse, nudging is sometimes equated or even confused with libertarian paternalism (Hansen, 2016). Some authors equate nudges with libertarian paternalism or at least restrict their discussion of nudges to those that have a paternalistic character (see for example Sunstein, 2014), while others see the two concepts as clearly distinct (Hansen, 2016). To address all criticisms in a systematic manner, we will dissect Thaler and Sunstein's definition into its separate elements and discuss them one by one. Subsequently, we will review elements that have been proposed as to be added to the definition. The Thaler and Sunstein (2008) definition contains the following elements:

1. any aspect
2. of the choice architecture
3. that alters people's behaviour
4. in a predictable way
5. without forbidding any options
6. without significantly changing economic incentives
7. that is easy and cheap to avoid

1) “Any aspect”. It is contested whether nudges are aspects of the choice architecture, as the word ‘aspect’ indicates a characteristic that is not necessarily created intentionally. If nudges are indeed defined as aspects, accidental properties would also count as nudges. Alternatively, it has been argued the definition of nudges should be restricted to intentionally created aspects (Hansen, 2016). Defining nudges as intentionally created aspects fits better with the idea that nudges are consciously created by an agent to achieve some goal.

2) “Of the choice architecture”. Choice architecture is understood as “the context in which people make decisions” (Thaler & Sunstein, 2008, p. 3). A somewhat more concrete definition of choice architecture is “the way in which choices are presented” (Bovens, 2009, p. 2). This formulation signifies that nudges aim to change the way in which choices are presented, not choices themselves, even though the distinction between a choice and the presentation of a choice is not always clear-cut.

3) “That alters people’s behaviour”; and 4) “In a predictable way”. These elements have not been contested in the literature. The function of these elements is to indicate that nudges aim to realize behavioural change in the desired direction of the nudger, which is broadly accepted as a necessary component of the definition of nudge.

5) “Without forbidding any options”. Saghai calls this element the ‘Choice-Set Preservation Condition’ and formalizes it in the following way: “A preserves B’s choice-set when the choice-set is unaltered or expanded, compared to a baseline representing B’s situation prior to A’s influence attempt” (Saghai, 2013, p. 488). The only ambiguity regarding this element of the definition is that it is unclear if the removal of irrelevant options from the choice-set counts as restricting the choice-set (Hansen, 2016). This is a relevant question because in marketing it is known that consumers can be influenced by adding irrelevant but expensive options to an array of choices because it makes the real options look cheaper by comparison. This mechanism is known as the decoy effect.

6) The phrase "without significantly changing economic incentives" (Thaler & Sunstein, 2008, p. 6) triggers the question why only *economic* incentives should be unchanged. Several authors have critiqued the exclusion of non-monetary incentives. Hansen and Jespersen (2013, p. 7) note that under Thaler and Sunstein's definition, giving a person electric shocks would count as a nudge. More realistically, Hausman and Welch ask if social pressure should be included, which is important as many nudges make use of social pressure (Hausman & Welch, 2010). Thaler and Sunstein have replied to this critique and state: "Some of our nudges do, in a sense, impose cognitive (rather than material) costs, and in that sense alter incentives. Nudges count as such [...] only if any costs are low" (Thaler & Sunstein, 2008, p. 8). This brings us to the question: what counts as significantly low? We will address this point under element 7.

7) "Easy and cheap to avoid". Thaler and Sunstein want to "allow people to go their own way" (Thaler & Sunstein, 2008, p. 249). This is a central tenet in their thinking, and the basis for calling libertarian paternalism *libertarian*. Some authors refer to this idea as 'reversibility' (Rebonato, 2014), to indicate that the influence nudges exert can be overturned by the nudgee. The effects of nudges are in that sense 'reversible'. This element is about securing a more substantial freedom of choice than simply maintaining the choice-set (element 5). Saghai (2013) refers this idea as "easy-resistibility" (p. 489) or "substantial-noncontrol" (p. 488). For him, "A's influence is easily resistible if B is able to effortlessly oppose the pressure to get her to ϕ if she does not want to ϕ" (Saghai, 2013, p. 489). The substantial noncontrol criterion thus excludes from the definition interventions that are not easily resistible. Some authors that exclude interventions that are substantially controlling while maintaining the choice-set from the definitions of nudge refer to such interventions as either "shoves" (Whyte et al., 2012, p. 35) or "prods" (Saghai, 2013, p. 491). While Saghai does not regard substantially controlling interventions as nudges, other authors have

criticized nudges for precisely this reason, namely that nudges have an illegitimately strong influence on people (see e.g., Hausman & Welch, 2010). It can be inferred that these authors do not take nudges to be substantially non-controlling by definition. Hence, they must either have some other criterion for "easily avoidable" or reject the idea that to count as a nudge an intervention must be easy and cheap to avoid altogether. The controversy over how easily resistible nudges are by definition, is one of the most prominent in the literature.

Having discussed all elements of Thaler and Sunstein's (2008) definition, we turn to elements that have been suggested as lacking from their definition. In the literature, two additional elements were found.

8) Shallow cognitive processes. It is argued that nudges target the "fast", "automatic" and "intuitive" system 1 processes in dual process theory. Several authors argue this element must be included in the definition of nudge. This element can be referred to in multiple ways. Saghai states nudges work "by making use of shallow cognitive processes" (Saghai, 2013, p. 489). Alternatively, Engelen and Nys formulate the same idea by stating nudges work "by tapping into a-rational psychological mechanisms" (Engelen & Nys, 2020, p. 138).

However, it is controversial if shallow cognitive processes must be targeted by an intervention for it to count as a nudge. Some authors point to the existence of nudges that instead of heuristics and biases, trigger reflective and deliberate processes (Acquisti et al., 2017). The question then is whether aspects of the choice architecture that foster critical thinking should be considered nudges.

Accordingly, some authors that include both cognitive processes in the definition of nudging distinguish 'type 1-' from 'type 2 nudges' (Hansen & Jespersen, 2013). Yet others formulate a distinction between 'heuristics-triggering-' and 'heuristics-blocking nudges' (Barton & Grüne-Yanoff, 2015, p. 343), which are also referred to as 'system 1-' and 'system 2 nudges', in line with dual process theory (Sunstein, 2016; Lin et al., 2017).

In contrast, authors who understand nudging to target automatic (system 1) processes exclusively have developed different concepts to refer to type 2 interventions. Hertwig and Grüne-Yanoff (2017) for example distinguish nudges from 'boosts', where boosts refer to type 2 interventions that aim to bring about behaviour change, as "their goal is either to train the functional processes or to adapt the external world (e.g., representation of information), or both, to improve decision making and its outcomes" (Hertwig & Grüne-Yanoff, 2017, p. 7).

These distinctions are important for the ethical discussion on nudging, as some authors connect this distinction to their respective normative appraisals (cf. Gigerenzer, 2015). However, Sims and Mueller (2019) argue no normative evaluations can be "read-off from type membership in this way" (p. 219).

9) Libertarian paternalist vs. technical definitions. Several authors include an element of benevolence in their definition of nudge, as noted by Hansen (2016). These authors argue that nudges are by definition interventions that aim to improve people's lives, often, but not always, by referring to Thaler and Sunstein's (2008) credo of "making people better off, as judged by themselves" (p. 5). Authors that incorporate an element of benevolence in their definition of nudge argue that the goal of a nudge is by definition to improve (some aspect) of welfare (see e.g., Sunstein, 2014; Brunon-Ernst, 2018). We will call definitions that include an element of benevolence libertarian paternalist definitions of nudge. In contrast, Nys and Engelen (2017) argue that "nudging refers to techniques used to influence people's behaviour" (p. 201), regardless of the motive to influence. We will call these definitions technical definitions of nudge. Most entries in the literature use a libertarian paternalist definition of nudges. However, it must be noted that many authors only implicitly subscribe to a libertarian paternalist view of nudges. These authors do not explicitly discuss welfare in their definition of nudge but mention exclusively normative examples

and make ethical arguments that refer only to the welfare-enhancing credentials of nudge.

A further controversy regarding the normative definition is *for whose* good nudges must be. On the one hand, it can be argued that nudges should be for the good of the nudgee (accepting Thaler and Sunstein's (2008, p. 5) making people "better off as judged by themselves"). This approach would include nudges to stimulate prudent saving. On the other hand, it can be argued that nudges should be for the good of society. This idea would include nudges that stimulate environmentally conscious choices. This distinction has been referred to as 'pro-self'- versus 'pro-social' nudges (Barton & Grüne-Yanoff, 2015), or alternatively as 'self-regarding'- versus 'other-regarding' nudges (Clavien, 2018; M'hamdi et al., 2017).

Alternative definitions of nudges

Given the critique of Thaler and Sunstein's (2008) definition, several authors have proposed an alternative definition. Hausman and Welch define nudges as: "ways of influencing choice without limiting the choice set or making alternatives appreciably costlier in terms of time, trouble, social sanctions, and so forth" (Hausman & Welch, 2010, p. 126).

Hansen and Jespersen propose a revision of Hausman and Welch's definition:

> A nudge is any attempt at influencing behaviour in a predictable way without forbidding any previously available courses of actions or making alternatives appreciably costlier in terms of time, trouble, social sanctions, and so forth. In other words, nudging is not necessarily about the manipulation of 'choice'. (Hansen & Jespersen, 2013, p. 15)

According to Saghai:

A nudges B when A makes it more likely that B will ϕ, primarily by triggering B's shallow cognitive processes, while A's influence preserves B's

choice-set and is substantially noncontrolling (i.e., preserves B's freedom of choice). (Saghai, 2013, p. 491)

Hansen proposes the following definition:

> A nudge is a function of (I) any attempt at influencing people's judgment, choice or behaviour in a predictable way (1) that is made possible because of cognitive boundaries, biases, routines and habits in individual and social decision-making posing barriers for people to perform rationally in their own declared self-interests and which (2) works by making use of those boundaries, biases, routines, and habits as integral parts of such attempts. [And] works independently of: (i) forbidding or adding any rationally relevant choice options, (ii) changing incentives, whether regarded in terms of time, trouble, social sanctions, economic and so forth, or (iii) the provision of factual information and rational argumentation. (Hansen, 2016, p. 158)

Engelen and Nys understand nudges as:

> deliberate changes in people's choice architectures with the intention of predictably influencing their behaviour by tapping into a-rational psychological mechanisms – often labelled 'heuristics and biases' – and thus without merely informing, rationally persuading, incentivizing, or coercing them. (Engelen & Nys, 2020, p. 138)

They add that "if an intervention is not easily resistible, it ceases to qualify as a nudge" (Engelen & Nys, 2020, p. 138).

Alternative definitions of libertarian paternalism

Saghai defines libertarian paternalism as "a justificatory strategy for a subset of nudges satisfying two conditions: (1) they are performed for the benefit of the nudgee; (2) they satisfy an informed desire welfare criterion" (Saghai, 2013, p. 488). Note that under Saghai's definition of libertarian paternalism, it is a justification for self-regarding nudges and not for other-regarding nudges.

This view is shared by Korobkin (2009), who calls the justificatory strategy for other-regarding nudges “Libertarian Welfarism” (p. 1653).

Rebonato (2014) defines libertarian paternalism as:

> the set of interventions aimed at overcoming the unavoidable cognitive biases and decisional inadequacies of an individual by exploiting them in such a way as to influence her decisions (in an easily reversible manner) towards choices that she herself would make if she had at her disposal unlimited time and information, and the analytic abilities of a rational decision-maker (more precisely, of Homo Economicus. (Rebonato, 2014, p. 359)

Saghai and Rebonato both stress that nudging is a technique to influence behaviour while preserving freedom of choice, while libertarian paternalism is the theory that justifies the use of that technique to improve welfare. This conceptual distinction has moral relevance, because it is possible to accept nudging as a technique, while condemning libertarian paternalism (see e.g., Gigerenzer, 2015).

While these two definitions, like Thaler and Sunstein’s, leave open who is the performer of libertarian paternalism, several authors limit the term to use by governments and public policy. These authors define libertarian paternalism rather as the “political-philosophical underpinnings of nudging” (Gumbert, 2019, p. 96) or the “normative justification for behavioural policy” (Häußermann, 2020, p. 45).

2.4 Ethical issue: Autonomy

One of the most prominently discussed ethical issues in the literature is the concern that nudges violate autonomy. 86% of the included records mentioned this issue (Figure 6). However, different authors have different conceptions of autonomy. To make clear which conceptions of autonomy are used

in the literature on the ethics of nudging, Vugts et al. (2020) conducted a literature review on 33 articles and found three conceptions of autonomy used in the debate: 1) freedom of choice, 2) agency and 3) self-constitution. We found no reason to divert from this trifold conceptualization, so we use it to structure our discussion of the issue of autonomy.

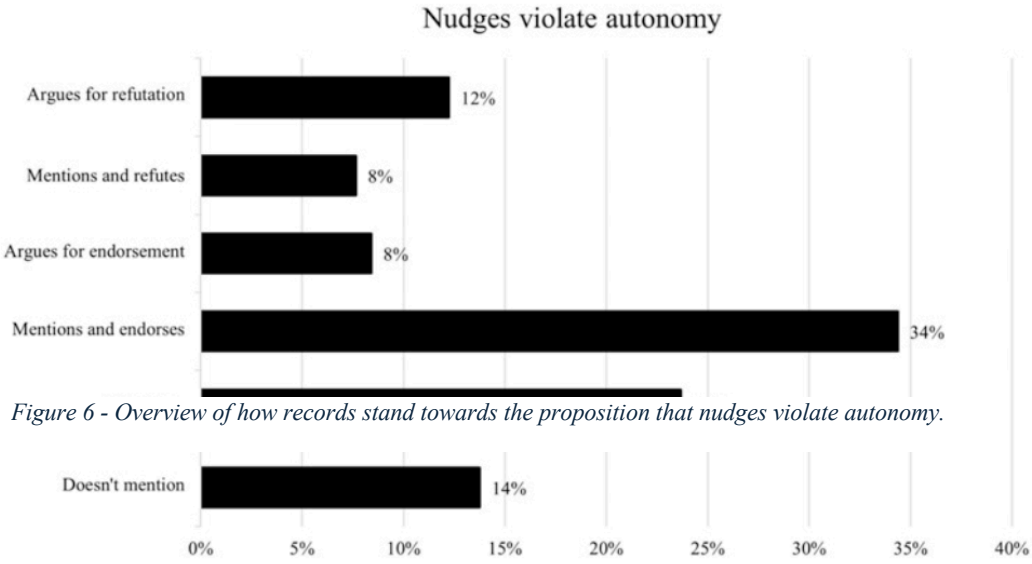


Figure 6 - Overview of how records stand towards the proposition that nudges violate autonomy.

2.4.1 Freedom of choice

Autonomy conceptualized as freedom of choice involves having access to relevant alternatives without being forced to make any particular choice (Vugts et al., 2020). Authors who emphasize the importance of freedom of choice use a negative conception of liberty as the absence of control, pressure, or coercion (Berlin, 2017). This form of autonomy encompasses the absence of external pressures on agents that might prevent them from acting as they wish. Thaler and Sunstein (2008) argue nudges respect autonomy understood as freedom of choice because nudges leave people's choice-set unchanged and do not provide significant (financial) incentives. Therefore, they argue, nudgees retains all options and are thus free to choose as they please: their autonomy understood as freedom of choice is respected.

In reviewing the literature, we found two strands of criticism that challenge the idea that nudges respect freedom of choice. The first strand of criticism argues that nudging violates the republican idea of freedom. Central in this ideal of freedom is the concept of arbitrary power, defined by Pettit as "(1) the capacity to interfere (2) with impunity and at will (3) in certain choices that the other is in a position to make." (Pettit, 1996, p. 578). Crucially, Pettit focuses on the capacity to interfere rather than actual interference. Grüne-Yanoff (2012) applies this argument to governmental nudges and libertarian paternalism as he states that governments can exert arbitrary power over its citizens using libertarian paternalistic policies, thereby violating their autonomy. Thus, the ethical issue that he raises with respect to nudges is that governments capable of nudging can shape their citizen's choices at will, which places them in a position of subjugation and constitutes a violation of autonomy. In the healthcare domain, a similar argument is made by Hamilton (2018) who uses the republican ideal of freedom to argue that physicians have arbitrary power over their patients when they nudge them to provide informed consent.

The second challenge to the idea that nudges respect freedom of choice is that nudges do not allow people to go their own way if they are not "easily resistible" (Saghai, 2013, p. 489). For Saghai, "[a]n influence does not sufficiently preserve freedom of choice if we are unable to easily resist it." (Saghai, 2013, p. 488). So, the easy resistibility criterion requires nudges not only to leave the choice-set intact, but also to give choosers a genuine opportunity to resist the influence of the nudge. If the force of the nudge is of such strength that it substantially impacts agent's ability to make their own decision, the nudge violates their autonomy understood as freedom of choice, even if it leaves their choice-set unaffected. This critique of nudging can be reformulated as the idea that nudges are either effective, or easily resistible but cannot be both. Illustratively, Floridi (2016) argues there is a trade-off between the effectiveness

and the easy resistibility of nudges. A similar argument can be found in Mills (2018).

Some authors aim to settle the question of whether nudges respect freedom of choice by empirical arguments that should prove the resistibility of nudges. Viale (2018) argues neurological processes prevent us from being cognitively able to resist the influence of nudges. However, this is contradicted by another empirical argument as Felsen and Reiner (2015) argue based on neurological evidence that nudges do not influence people more strongly or more invisibly than other factors that people encounter in everyday life. They argue that because of this, nudges do not hurt autonomy in any relevant sense.

2.4.2 Agency

Autonomy as agency involves capacity to reason, critically reflect, and make choices (Vugts et al., 2020). In contrast to autonomy as the absence of external pressures, this is a more positive conception (Berlin, 2017). To be autonomous requires not only having options and being free *from* interference but also having the internal capacity to reflect on the options provided and being free *to* act on them to achieve personal goals. A further difference with the first conception of autonomy is that this conception concerns the internal psychological processes of agents, as it requires mental capacities to deliberate, critically reflect and reason about preferences.

Nudges are criticized for violating autonomy as agency because they target shallow cognitive processes - the fast, automatic, and intuitive type 1 processes - to alter people's behaviour (Saghai, 2013). In this way, it is argued, nudges fail to treat people as reasoning subject and so diminish control and agency (Hausman & Welch, 2010; Lanzing, 2019; Noggle, 2018). Bovens argues: "there is something less than fully autonomous about the patterns of decision-making that Nudge taps into. [...] we are not fully in control of our actions" (Bovens, 2009, p. 4). An example of such an argument can be found in MacKay and Robinson (2016) who focus on a specific type of nudges, namely defaults. When

presented with a choice defaults are the alternative which the chooser “will obtain if the chooser does nothing” (Thaler & Sunstein, 2008, p. 83). For example, in certain countries citizens are considered organ donors by default whereas in other countries, they are not considered organ donors by default. Setting a default is a nudge because it does not affect the choice yet has an effect on which alternative is chosen by most users. This can be explained by inertia, or by choosers experiencing the default as a recommendation. MacKay and Robinson argue that defaults seek to circumvent rather than engage reason and argumentative capacities. In their view, defaults circumvent rather than engage a reflective and deliberative choice and therefore harm autonomy as agency. In sum, autonomy as agency requires that agents are in control over their decisions by reflecting and deliberating about preferences. Nudges are critiqued for diminishing or circumventing critical reflection. This could have the effect that people make choices they might never make out of their own accord (Muldoon, 2018).

Engelen (2019a) refers to the idea that agents are rational when they are able to think and reflect clearly as "process-rationality" (p. 218). Being rational is then to be able to reason and reflect on preferences and beliefs. However, Engelen argues that people should be called rational when they attain their desired outcomes, irrespective of the mental processes that they use to achieve these outcomes. He refers to his form of rationality as "outcome-rationality" (Engelen, 2019a, p. 222). Outcome rationality differs from process rationality in that it does not require agents to think rationally or deliberate about their choices. It only requires they make choices that are in line with their core values.

Engelen argues that it is rational for agents to act in a way that allows them to realize their self-chosen goals. Nudges undermine agents' rational thinking by targeting shallow cognitive processes. However, they can steer towards choices the agents endorse and allow them to resist those nudges towards choices they do not endorse. Ergo, a nudge is compatible with their

status as rational agents. This argument illustrates that rationality is not necessarily part of autonomy as agency, and that it is disputed whether the irrational influence that nudges exert is a sufficient basis to conclude nudges violate autonomy.

An important and often mentioned sub-worry regarding agency is that of manipulation. Manipulation is considered to be intentionally subverting someone's deliberative capacities in a surreptitious way (Wilkinson, 2013). A such, manipulation prevents agents from reasoning about their preferences and is therefore a form of harming someone's autonomy as agency. Manipulation is conceptualized as circumventing people's "capacity for reason [...] by exploiting nonrational elements of psychological makeup" (Blumenthal-Barby & Burroughs, 2012, p. 5), or as "using methods that pervert choice" (Wilkinson, 2013, p. 347). The manipulation critique thus states that nudges target what was previously defined as shallow cognitive processes and thereby trick agents into making decisions they would never have made on their own.

However, to call a nudge manipulative it must also be surreptitious; it must influence the nudgee in a secret, opaque way. This is because when agents discover they are being manipulated, they can correct the manipulative influence and restore their agency (Bovens, 2009). The surreptitious component of manipulation is voiced by several authors with respect to nudges: they are a form of "hidden steering" (Lanzing, 2019, p. 553), they occur "in a way that is not obvious to the subject" (Blumenthal-Barby and Burroughs, 2012, p. 5) and "without making it apparent that this is occurring." (Aggarwal et al., 2014, p. 3). Accordingly, it has been argued that nudges are ethical, if they are transparent (Kanev & Terziev, 2017). For example, Bovens (2009) argues that to avoid manipulation, every nudge should in principle be detectable for a watchful person. Hansen and Jespersen argue transparency can even be considered as the main criterion to evaluate whether a nudge is manipulative (Hansen & Jespersen, 2013). The problem here is that the covert nature of nudges usually determines

their success; both Bovens (2009) and Grüne-Yanoff (2012) argue nudges will be more effective when they are opaque. An often-cited phrase is that “nudges work best in the dark” (Bovens, 2009, p. 13). Hence, it might be difficult to nudge both effectively and transparently. To reduce the tension between the two, Ivankovic and Engelen (2019) offer a new concept: watchfulness. To be watchful is the “capacity of individuals enhanced by resources at their disposal to detect ex post and non-transparent nudges that undermine their aims” (Ivankovic and Engelen, 2019, p. 23). They also provide ways in which societies can ensure they are sufficiently watchful of nudges.

There is controversy over the relationship between manipulation and transparency. Some reject that transparency is sufficient for acceptable nudging because “[b]eing open about nudges does not make them less manipulative.” (Nys and Engelen, 2017, p. 205). For these authors, the manipulation is to influence agents’ choices so that they fail understand the reasons for their own behaviour. Whether the failure to understand themselves is caused in a transparent or in an opaque way is irrelevant. In this sense, these authors argue, nudging techniques are manipulative, even when the nudge is transparent. A similar argument is advanced by Glod (2015), who argues that the fact that nudges aim to target shallow cognitive processes, rather than that they are opaque, is what makes them manipulative.

2.4.3 Self-constitution

The third conception of autonomy is concerned with identity, individuality, authorship, and authenticity (Vugts et al., 2020). It emphasizes that one's "ultimate values are not contingent; they are essentially his or hers – the result of a reflective endorsement that constitutes his- or herself" (p. 11). When autonomy as self-constitution is violated, people may still have significant attainable choices (freedom of choice) and may still be able to reason about how these options relate to their preferences (autonomy as agency), but their deepest

values and preferences are not truly their own. A violation of autonomy as self-constitution is thus the negation of someone's authorship over their own life, prevention of someone's ability to set their own ends or the imposition of alien values.

As autonomy as self-constitution requires that agents can set their own ends, it requires that agents do not simply have preferences over choice alternatives, but that they also have higher-order preferences over their lower-order preferences. As this conception of autonomy presumes a hierarchy between preferences, it is sometimes called hierarchical autonomy, which is often attributed to Harry Frankfurt (Blumenthal-Barby, 2010).

Nudges harm autonomy as self-constitution when agents' lower- and higher-order preferences are aligned, and a nudge persuades them to act contrary to their higher-order preferences. For example, when someone's aim is to eat healthily, yet is nudged in the supermarket to buy junk food. In such a case, a nudge can lead to a situation in which agents do not understand their own behaviour as they are nudged to act contrary to their higher-order preferences (Bovens, 2009). This concern is central to the issues of autonomy as self-constitution. Iyer (2017) adds to this discussion from a virtue ethics lens and argues it is important to act for the right reason. If a nudge is the reason for acting, the actors do not engage, nor train their practical intelligence.

However, Nys and Engelen (2017) argue it is unlikely this will happen: “[t]he more a nudge goes against a person’s settled preferences, the more it makes her aware of the influence (ceteris paribus)” (p. 204). If agents are nudged towards a choice they endorse, the nudge is seamless, and agents will intuitively follow it. If on the other hand they are nudged towards a choice they strongly dislike, they will become aware of the influence, making the nudge transparent to them, and they will become able to resist it. Thus, according to Nys and Engelen, a nudge is effective for those who endorse the nudged

alternative, yet resistible for those who reject it and hence capable of respecting autonomy as self-constitution.

Another argument in defence of nudging states that in situations where higher- and lower-order preferences diverge, a nudge could restore unity by influencing agents to behave in line with their higher-order preferences (Bovens, 2009). Hierarchy of preferences can explain situations of akrasia, in which an agent's a higher-order desire to abstain from a hedonistic pleasure is overruled by a lower-order desire to indulgence in that pleasure. It has been argued that nudges can help to choose in accordance with higher-order preferences and so increase the agent's autonomy (Engelen & Nys, 2020). Barton (2013) argues that unwilling smokers can be made more autonomous by nudging them away from smoking because it improves their ability for self-rule. Sawicki states that the irrational influence of nudges can still enhance one's autonomy if the nudge helps to override unwanted "internal or external hindrances" (Sawicki, 2016, p. 223). Engelen (2019a) argues that in those situations, process rationality (agency) is sacrificed for outcome rationality (self-constitution). So, while nudges are means of irrational persuasion, they can help to attain rational outcomes when they bring in line our lower-order with our higher-order preferences.

Critics nevertheless argue that nudges do not merely trick agents to choose in conflict with their higher-order preferences but corrupt people's higher-order preferences. Hence, it is argued that nudges violate autonomy as self-constitution because they interfere with agents' ability to choose their own ends and preferences. This issue is raised by Baldwin (2014) who argues that nudges "involve re-shaping the target's idea of their own welfare" (p. 14). A closely related critique is that nudges may lead to a "fragmented self" (Bovens, 2009, p. 8), which means that if nudges successfully exert influence on someone and they make choices they would never have made of their own accord, they may no longer recognize their choices as their own. When nudges often succeed in changing behaviour, people might come to question their true preferences and

even start to doubt their ability to form proper judgment (Dold & Schubert, 2018).

Libertarian paternalism and normative nudges are especially vigorously criticized for violating autonomy as self-constitution because these nudges imply that some authority identifies bad choices and tries to replace them with better alternatives. Some authors interpret this as disrespectful as it "suggests the superiority of those who design the nudges or perhaps even their contempt for those whom they nudge" (Hausman, 2018, p. 277). The ethical issue that nudges violate autonomy as self-constitution thus includes the critique that nudges show a lack of respect. This lack of respect critique has been alternatively formulated as: "disrespectful social control" (Hausman & Welch, 2010: 134), "a violation of dignity" (Mortensen et al., 2019), "contempt" (Hausman, 2018, p. 277), "stigmatizing" (Hukkinen, 2016), "insulting" (Kasperbauer, 2017, p. 54), "condescend[ing]" (Jennings et al., 2016, p. 92), "reducing the citizen's voice in his or her destiny" (Baldwin, 2014, p. 19), nudges "treat people like children" (Hausman, 2018, p. 276) and treating people as "puppets on strings" (Wilkinson, 2013, p. 342). The common denominator of these critiques is that nudging insufficiently recognizes people as a source of authority over their own preferences and thus violate autonomy as self-constitution.

2.5 Ethical issue: Welfare

The issue of welfare is concerned with the proposition that nudges improve welfare. This is mostly in accordance with or in response to Thaler and Sunstein (2008), who argue governments should nudge citizens to ensure they make welfare-optimal decisions. They argue this is necessary as many people frequently act irrationally. Irrational behaviour is here understood as acting against one's own interests. To illustrate the ubiquity of irrational behaviour, Thaler and Sunstein (2008) list many examples in which people make choices that fail to optimize their welfare. The two most prominent examples of

irrational choices are 1) people who choose unhealthy food while claiming to be on a diet; and 2) people who maintain credit card debt against expensive interest rates while having money in their saving accounts. Therefore, Thaler and Sunstein argue that governments should not assume that people behave rationally but instead take note of the ways in which people frequently behave irrationally and help them to make better choices to optimise their welfare.

Importantly, Thaler and Sunstein do not presuppose any idea of what constitutes welfare (Thaler & Sunstein, 2008). People should be nudged to choose the things that maximise their welfare in line with their own individual conception. Thaler and Sunstein's idea of welfare has been called an "informed preference" criterion of welfare (Rebonato, 2014, p. 362), because it seeks to steer people toward the choices that they would make themselves if they were completely rational and informed. This is reflected in Thaler and Sunstein's adage: making choosers better off "*as judged by themselves*" (Thaler & Sunstein, 2008, p. 5). Thaler and Sunstein's argument has found agreement in the literature on nudging (see Figure 6). As stated in the definition section, for many authors nudges are equivalent with the goal of improving the welfare of the nudgees. We will refer to the authors that argue nudges should be used to improve welfare as libertarian paternalists. However, as discussed, it is disputed that nudges always improve welfare. In this section on the issue of welfare, we distinguish eight arguments against libertarian paternalist nudges and the idea that they can improve welfare.

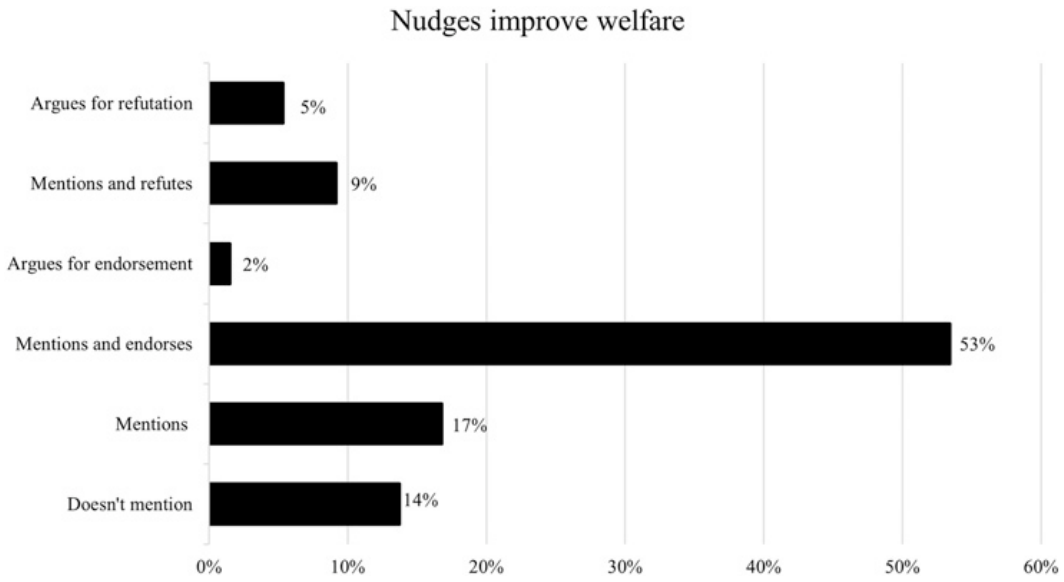


Figure 7 - Overview of how records stand towards the proposition that nudges violate improve welfare.

2.5.1 Existence of true preferences

The first argument against authors who argue nudges improve welfare is that their conception of welfare presupposes the existence of true preferences while true preferences do not exist. Thaler and Sunstein's argument for welfare improving nudges presupposes the existence of true preferences because their argument to nudge people towards better choices needs a direction to nudge people into. According to Thaler and Sunstein, the optimal direction is towards welfare maximisation. When they state an agent makes an error, and that choices can be optimised in terms of welfare, they presuppose the possibility of ranking choices based on their expected utility. They refer to the alternative with the highest expected utility as an agent's "true preferences" (Thaler & Sunstein, 2008, p. 2). The underlying premise of the argument is thus that a set of optimal preferences exists for every individual. For proponents of libertarian paternalist nudges, the fulfilment of true preferences functions as a normative criterion. This welfare criterion is the key justification for interfering with people's choices in the nudge debate. However, the existence of true preferences is disputed. It is argued that even when people show inconsistent and contradictory preferences, none of

the preferences reflect the person's *true* preferences (Sugden, 2008; Lecouteux, 2015). Critics argue that proponents of libertarian paternalist nudges lack a way to assess an individual's true preferences (Glod, 2015). Libertarian paternalists reply that true preferences are those preferences that individuals would have if they were completely informed and had unlimited cognitive abilities. They thus reconstruct what agents would prefer in ideal circumstances. Critics find this answer unsatisfactory; they argue that such a reconstruction of preferences is implausible and cannot be performed objectively. Sugden asks: "Even if we can specify what it would mean to have these supernatural powers, how do we discover how some ordinary human being would act if he were somehow to acquire them?" (Sugden, 2008, p. 232). Therefore, the reconstruction of true preferences is criticized as "preference purification" (Häußermann, 2020, p. 47) or preference "laundering" (Dold & Schubert 2018, p. 226). The true preferences critique of welfare promoting nudges is thus that since it is impossible to reconstruct someone's true preferences, true preferences cannot be used as a justification for nudging. Critics who dispute the possibility of reconstructing true preferences argue that every attempt at reconstruction is an imposition of one's own values. Blumenthal-Barby argues that the choice architect's assessment of what others truly prefer amounts to a "perfectionist account of well-being" (2010, p. 2). Wright and Ginsburg add that nudges often favour "long-run ex ante preferences" over more short-term hedonistic in-the-moment preferences without "a defensible basis" (Wright & Ginsburg, 2012, p. 1061). Glod argues that instead of aiming to fulfil an informed-desire criterion of welfare, nudgers impose their own idea of what constitutes wellbeing on others, giving insufficient weight to the particular situations and views of different people (Glod, 2015). For Grüne-Yanoff, this means libertarian paternalism is unable to incorporate "the plurality nor the subjectivity of people's values" (Grüne-Yanoff, 2012, p. 644). Hence, the critique is that libertarian paternalists try to nudge people towards

choices that are healthy and financially prudent while there is no way to determine whether these goals are genuinely and truly preferred by all agents. Without the notion of true preferences, a sound justification for nudging people towards these choices, which they would not have made of their own accord, is lacking. As a result, it has been argued that the proponents of libertarian paternalist nudges and particularly Thaler and Sunstein, "fail to take seriously their targets' true preferences" (Scoccia, 2019, p. 79).

2.5.2 Irrationality leading to suboptimal choices

For proponents of welfare nudges, the reason that agents fail to optimise their welfare is that due to bounded rationality, willpower and self-interest, agents fail to make optimal choices. Agents suffer from "limitations of attention, information, cognitive ability, or self-control" (Häußermann, 2020, p. 47) and hence fail to identify which choices will lead to the greatest welfare. Instead of reasoning about their preferences, agents rely on fast heuristics, which make them susceptible to biases such as hyperbolic discounting, status quo bias, endowment effects, hot and cold states, social proof, and optimism- and availability bias (Kahneman, 2011). The result is that people's preferences are inconsistent: they spend money on diet programs while eating too much (Glod, 2015). They maintain saving accounts while paying for expensive credit (Thaler, 2015). The key conviction in a large part of the literature on nudging is that "inconsistent choices may be treated as mistakes" (Häußermann, 2020, p. 47).

Two critiques have been proposed against the idea that people are irrational. First, it is argued that inconsistent choices are not as common as claimed. Rebonato argues that many findings on behavioural mistakes are "artfully overemphasized in their literature" (Rebonato 2014, p. 392). In a similar vein, many economists criticize the behavioural findings for being done in laboratory settings that fail to mimic 'the real world' (Lodge & Wegrich, 2016) and that many findings of irrationality are "mere artefacts of the experimental

design" (Barton & Grüne-Yanoff, 2015, p. 353). Relatedly, it is argued that behavioural experiments usually involve small incentives, suggesting that people make better decisions when the stakes are higher (Glaeser, 2005).

Second, it is argued that proponents of libertarian paternalist nudges are too quick to dismiss choices as mistakes (Baldwin, 2014). For example, Glod offers an example about people who eat unhealthily, while saying they need to lose weight. He argues that it might be the case that this person does not act contrary to their preferences but rather engages in "cheap talk" (Glod, 2015, pp 616-617) suggesting people genuinely prefer short-term pleasure (smoking, fast food, spending money) while claiming to prefer long-term prudent choices that are the norm. Governments should acknowledge citizens enjoy such pleasures. Likewise, it is argued that in healthcare, physicians must admit that not all people value health as strong as they do. For these people, it might not be a mistake to refuse pre-emptive screenings or vaccinations (Ploug et al., 2014; Giubilini et al., 2019).

Critics of libertarian paternalist nudges thus state that people who prefer short-term pleasure over long-term health and wealth are too eagerly judged to make mistakes. Libertarian paternalists are too focused on the aforementioned values, health, wealth, and happiness, and are therefore too willing to argue that "departures from those preferences represent defects" (Wright & Ginsburg, 2012, p. 1061). Consequently, the argument is that libertarian paternalists are overly eager to dismiss choices they do not endorse as mistakes and are quick to dismiss alternative ideas of what constitutes a good life.

2.5.3 Epistemic access

Critics of libertarian paternalist nudges argue that even if true preferences would exist, and even if people were irrational, it would still be impossible to effectively nudge people towards more welfare because it is impossible to uncover people's true preferences, due to limited epistemological access to these preferences (White, 2013). It is argued that it is impossible for the government

to assess properly which nudges would optimise citizen welfare. To make this argument Rizzo and Whitman quote Bentham: “It is a standing topic of complaint, that a man knows too little of himself. Be it so: but is it so certain that the legislator must know more?” (Bentham, 1996, XVII.17). The argument that governments cannot know what citizens truly prefer was prominently made by Hayek (Binder, 2014). The Hayekian knowledge problem states that the knowledge in a well-functioning economy is so dispersed and decentralized that it would be impossible to centrally gain sufficient knowledge to manage the economy effectively. This argument is especially forceful because of libertarian paternalist’s insistence that nudges should make everyone better off on their own terms. The key idea of the criticism is that people are the best judges of what brings them happiness, even if they sometimes err. The government has no meaningful access to this knowledge. Therefore, this argument is referred to as the argument of constrained epistemic access (Fumagalli, 2016).

2.5.4 Heterogeneity

Thaler and Sunstein (2008) insist everyone should be made better off on their own terms. However, it is argued that in a population of heterogeneous people, a one-size fits all nudge steers some people away from well-being (Korobkin, 2009; White, 2017; 2019). The worry is that due to the large differences within population, a part of the population is nudged into directions they disapprove of (Engelen, 2019a). Binder (2014) connects the heterogeneity argument to the earlier critique that nudges steer people towards general benefits such as health and wealth, rather than towards their own true interests. He argues that libertarian paternalism misses the heterogeneity point because they wrongly assume everyone equally values these typical goods.

This criticism is further extended by Rizzo and Whitman (2009a), who argue that to construct effective nudges, governments must not only know their citizen’s true preferences, but must account for the fact that heterogeneity has

multiple dimensions (Rizzo & Whitman, 2009a). Firstly, nudgers would have to account for pluralism, as different people have various conceptions of the good (Grüne-Yanoff, 2012; Ewert, 2017). Secondly, nudgers would have to account for differences in behaviour. Whereas some people save too little, some may save too much. A one-size-fits-all nudge to promote saving would then push the latter group in the wrong direction. Thirdly, biases are not uniform among the population and people may have different self-correction methods (Rizzo & Whitman, 2009a). To effectively nudge an agent to his optimal welfare, not only knowledge of their preferences is required, but also knowledge of how far they are away from their optimal choice to devise the optimal nudge. Fourthly, nudgers would have to account for differences in how people react to a nudge (Baldwin, 2014). Finally, reactions to one specific nudge may differ especially between social, economic, and cultural groups (Selinger & Whyte, 2011).

2.5.5 Privacy

In response to the heterogeneity argument which states that if nudges seek to optimize individual welfare, they cannot rely on a one-size-fits-all approach, Sunstein (2016) argues nudges should be personalized to nudge every agent towards their personally optimal alternative. According to Sunstein, especially defaults are suited for personalization. The idea is simple: if agents must make a complicated choice with many alternatives, for example choose a healthcare plan among many options, they could be nudged by preselecting a personalized default option for them, based on their characteristics such as age, income, and medical situation or based on behavioural data. In this way, personalized nudges can circumvent the heterogeneity problem.

Various authors have reacted to personalized nudges and argue this raises a new worry regarding privacy (Acquisti et al., 2017; Nahmias et al., 2019). Critics worry that personalized nudges demand the collection of citizen information on a gigantic scale, leading to the violation of privacy. Kapsner and Sandfuchs argue

that the need for information about preferences is “the source of one of the greatest temptations for libertarian paternalists when it comes to collecting information about citizens and invading their private spheres.” (Kapsner & Sandfuchs, 2015, p. 458). In addition, the large-scale collection of citizen data is criticized on its potential for misuse. Ranchordás (2019) argues data are mostly processed automatically, leading to algorithmic profiling risks. While this argument applies solely to governmental nudging, the privacy worry also arises in business. With the notable difference that in business, the goal of collecting data is often not to find out which nudges optimize individual welfare, but to discover which nudges are most effective at persuading customers to buy or use products and services. Such nudges likewise produce threats to privacy (Smith et al., 2013) and of algorithmic bias (Lanzing, 2019).

Online nudges that are personalized and optimized for effectiveness with the help of user data have been called “hypernudges” (Yeung, 2017, p. 118), “digital nudges” (Karlsen and Andersen, 2019, p. 1) or “data-driven nudges” (Ranchordás, 2019, p. 12). With these digital nudges, it is easy to collect statistics, giving choice architects and their algorithms the possibility to learn which nudges are most effective for each type of user. This makes such digital – personalized – nudges potentially more effective than their offline counterparts, hence the name ‘hypernudge’ (Yeung, 2017).

2.5.6 Cognitively limited nudger

The rationale for libertarian paternalist nudging is that people make sub-optimal choices due to cognitive biases and errors. It is argued that libertarian paternalist nudging will be unsuccessful because choice architects are equally susceptible to biases and errors as the people they seek to nudge (Blumenthal-Barby, 2010). The argument of the cognitively limited nudger thus states that it is impossible for the government to help people make better choices, since bureaucrats suffer from the very same cognitive biases.

Lodge and Wegrich (2016, p. 250) call this the "rationality paradox of nudge", as libertarian paternalism stresses that people are biased but is blind to how irrationality affects the execution of libertarian paternalistic policies. In other words, based on evidence that all people suffer from cognitive biases, it is incorrectly argued we must grant a small number of people the power to make important choices on behalf of everyone. Rebonato calls the same idea the "libertarian paternalistic predicament" (Rebonato 2014, p. 379) and Binder (2014) refers to roughly the same idea as the 'compounded knowledge objection'. These problems are increasingly receiving attention in the literature. There is an emerging field of research called 'behavioural political economy' which investigates how behavioural ideas influence politics and the public domain (Schubert, 2017b).

2.5.7 Abuse and public choice problems

Another argument against the case for welfare-optimising nudges is that governments could abuse nudging or be either inclined or under pressure to use nudging for other goals than optimizing welfare (Calo, 2014; John, 2018). The literature on public choice investigates which incentives policymakers have. Arguments from this field are used to question "whether policymakers have the right incentives to implement wise policies, given their own self-interest and the lobbying efforts of interested parties" (Rizzo & Whitman, 2009a, p. 908). Various authors refer to Glaeser, who argues that paternalism has historically "been abused by governments responding to special interests or to seek to aggrandize their own authority" (Glaeser, 2005, p. 2). This critique is applied to libertarian paternalist nudges, of which it is argued that they are even more prone to abuse than traditional paternalist policies as they operate covertly, making it harder to detect abusive practices (Hausman & Welch, 2010). Several authors mention worries concerning public officials misusing the power of nudges, see Thaler and Sunstein (2008), Blumenthal-Barby (2010), and Wolcott (2019). The fear is that

public officials and politicians could employ nudges to organize the support they need for political goals (Calo, 2014). Worse still, Schubert (2017b) argues that policy makers have a long-term interest in aggravating thinking biases in the population as policy makers have an interest in a docile and following citizenship. He suggests that if this is the case, public officials might actively try to foster biases.

2.5.8 Exploitation of error and economic harm

Similarly, private actors might seek the exploitation of error. In the literature, there are worries concerning marketeers who seek to foster the biases of consumers to improve their profits (Glaeser, 2005). Marketeers use nudge techniques to attract consumer attention and persuade them into purchasing (Yeung, 2017). Consequently, consumers are nudged into buying things they do not need or even want (Smith et al., 2013). Companies thus "use intentional choice architecture to nudge us towards purchase decisions that are good for their profit line but in some ways bad for" consumers (Schmidt & Engelen, 2020, p. 7). In this way, nudge techniques can divert consumers, leading to suboptimal choices. Ultimately, if private parties succeed in massively inducing or fostering thinking errors, this could lead to inefficient markets in which consumers make poor choices, generating welfare losses and economic harm.

In reaction, the critique on private use of nudging techniques has been used to justify the use of governmental nudges, as it is argued that the government should counter private influences to protect consumers from adverse influences (Curchin, 2017). Bovens states that "the government may be fighting a losing battle" against private parties if it cannot use the same techniques (Bovens, 2009, p. 15). Even more pronounced, Schmidt and Engelen (2020) refer to the use of nudges by private parties; they claim the defining question in the debate on ethics of nudging is: by who we want to be influenced: private parties, or the government? In contrast, Baldwin (2014) and Gigerenzer

(2015) both object to the idea that governments should engage in a combat with private parties for control over relevant choice architecture. Both authors note that private parties are likely superior at constructing nudges, considering their ability to spend massive amounts on advertising. Alternatively, the government could deploy tactics to mitigate the influences of private choice architecture. Schmidt (2019) identifies five alternative strategies. First the government can "replace a private nudge with a public one" (p. 531). For example, they could ban cookie consent pop-ups on private websites, which generally nudge users toward accepting all cookies and replace them with a standardized pop-up. Second, the government can instigate regulation for which nudging techniques private parties are allowed to use. Third, the government can mandate that private parties nudge towards "good" choices. This is done in the food industry, where producers are compelled to nudge consumers to purchase healthier products. Examples are New York's attempted ban on large soda containers, and government mandated use of traffic light food labelling (Thorndike et al., 2014). Fourth, the government can aim to counteract private nudges by making them less effective. Fifth and finally, the government can seek a softer form of regulation by cooperating with private parties and aiming to convince them to abandon bad nudging practices (Schmidt, 2019). Lavi (2018) adds one more alternative and argues that the designers of "bad" or "evil" (p. 1) nudges might be recognized as being liable for the law. Gigerenzer (2015) adds to this list of options by arguing that the government should not seek to gain control over relevant choice architecture but should rather educate citizens to make them more resistant to the influences of choice architecture. The government could consider boosting citizens to promote their welfare (Grüne-Yanoff, 2018; Hertwig & Grüne-Yanoff, 2017). So, if the intention of the government is to counteract the influence of private nudges, designing public nudges is but one of their available alternatives. However, Sims and Mueller (2019) disagree that boosting would always be normatively superior

to nudging and challenge the idea that the distinction between nudging and boosting has a normative status.

2.5.9 Easy resistibility, transaction costs, and harm

Proponents of libertarian paternalist nudging argue that when agents are nudged into the wrong direction, they could simply resist the nudge without harm caused. So, the argument goes, the welfare improvement for some nudgees does not come at a cost for others. For this to be the case, two conditions need to be satisfied. The first is the condition of easy resistibility, which states that agents who do not wish to be nudged can disregard the influence of the nudge without significant effort. The second is the condition of no additional burden, which states that agents are not negatively affected by the nudge through additional effort or cost. If a nudge steers part of the population in a welfare promoting direction and at the same time avoids affecting all others adversely in any way, the nudge constitutes a Pareto-improvement, which means that the nudge improves total welfare without making anyone worse off.

As we have already addressed the easy-resistibility criterion in the definition and autonomy sections, we will now turn to the condition of no additional burden. Critics worry that nudges do in fact lead to an additional burden for people (Vallgårda, 2012). This argument states that nudges impose costs on those who do not prefer the choices they are nudged into. Nudges make one alternative easier and more intuitive, while making other alternatives more difficult to choose. Thaler and Sunstein note that some nudges "in a sense, impose cognitive (rather than material) costs, and in that sense alter incentives" (Thaler & Sunstein, 2008, p. 8). According to them, these costs must be small as nudges must still be "easy and cheap to avoid" (Thaler & Sunstein, 2008, p. 6). Because the costs of nudges are supposed to be small, proponents of nudging argue these costs do not significantly impact welfare considerations.

This is disputed by critics of nudging who argue the costs of nudging can be substantial and meaningful for those whose preferences diverge from the nudge. The costs of nudges can take various forms. To avoid a nudge, a person "may have to [1] give up time, [2] take on an annoyance, [3] walk a little further, [4] pay a little more money or forfeit some money or material goods, [5] execute willpower, [6] feel bad about himself (self-scorn), [7] undergo social scorn, or [8] have to think a little harder" (Blumenthal-Barby, 2010, p. 20). All these hindrances may be argued to incur a cost on nudgees. The first four hindrances describe how nudges make it more difficult or costly to for choosers to obtain their chosen alternative when they wish to resist a nudge. These costs have been described as "transaction costs" (Glod, 2015, p. 601) and "friction costs" (Mills, 2018, p. 404). The last four hindrances describe the mental, emotional, and social pressures that nudges can exert over people. Authors who want to emphasize these forms of cognitive costs have described them as a "psychological tax" (Kirchgässner, 2014, p. 12) and as an "aggregate burden" (Baldwin, 2014, p. 16). Some authors worry that libertarian paternalistic policies affect everyone, helping those who would make supposedly bad choices while hindering those who already make supposedly good choices (Baldwin, 2014). Baldwin argues this could be seen as unfair to well behaving people, as libertarian paternalism presents them with an additional burden with the sole aim of helping those behaving "less responsibly" (p. 16).

2.6 Ethical issue: Long-term adverse effects

The issue of long-term adverse effects is concerned with the worry that nudges create adverse effects that only become apparent over longer periods of time when people have encountered multiple instances of nudges. In the nudge literature, most authors believe the moral evaluation of nudges should proceed on a case-by-case basis (Sunstein, 2014, 2015; Hansen et al., 2016; Engelen, 2019a,

2019b; Houk, 2019; Schmidt & Engelen, 2020, Yeung, 2012). Most of the records in the literature review discuss the ethics of nudging without reference to adverse long-term effects (see Figure 8).

However, some authors disagree with such an approach and argue that nudges should be evaluated through a more dynamic lens (Binder, 2014; Coggon, 2020; Lepenies & Malecka, 2015). Binder and Lades criticize libertarian paternalism for its "mainly static outlook" because they argue the case for libertarian paternalism has "strong intuitive appeal when described as one-shot situations without any regard for intertemporal dynamics" (Binder & Lades, 2015, p. 17). It is their suggestion that some prominent ethical problems with nudging only become apparent when analysing their long-term effects.

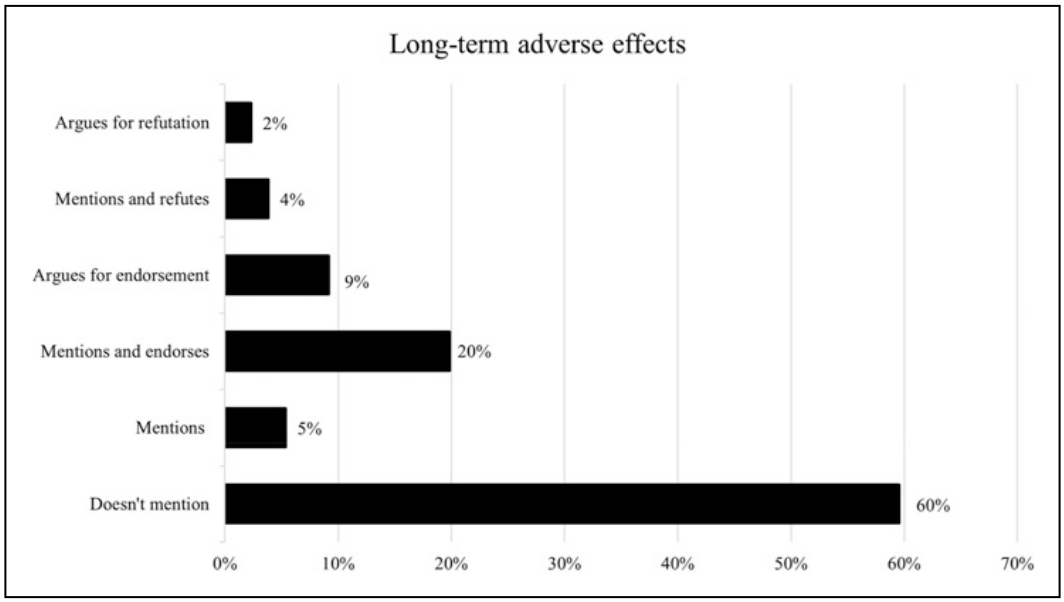


Figure 8 - Overview of how records stand towards the proposition that nudges cause long-term adverse effects.

2.6.1 Infantilization

The ethical issue of infantilization states that agents who are subject to many nudges over a longer period will develop eroded autonomous decision-making capabilities and will lack "responsibility in matters regarding [their] own welfare" (Bovens, 2009, p. 11). The infantilization critique is thought to be

especially forceful in discussions regarding libertarian paternalism, in which it is argued governments should widely adopt behavioural policies that nudge citizens towards better choices, as a great number of them will make people reliant on nudges. Schmidt refers to a similar idea as "the undermining worry" (2019, p. 536) and Riley (2017) uses the concept 'epistemic injustice'. Similar worries have been voiced by Binder and Lades, who state that "individuals may learn to be dependent and inactive" (Binder & Lades, 2015, p. 18), and Rebonato worries that in a nudged world, people are no longer stimulated to make deliberated decisions (Rebonato 2012). Moreover, one aspect of the infantilization issue is that nudging distorts an agents' process of learning what they prefer (Dold & Schubert, 2018) and hence lead to the "learning of faulty preferences" (Binder, 2014, p. 524). Various authors have argued that citizens should not be nudged to welfare but should rather be left alone so that they can experiment with their choices, so they learn what they truly value (Schubert, 2017b). Expanding on this point, Dold and Schubert (2018) argue that the process of learning what one's preferences are, is valuable in itself, as it allows agents to recognize their preferences as their own. Finally, Burgess argues that nudges impair agent's ability to improve morally (Burgess, 2012). In response to the issue of infantilization, several points have been raised. First, it is argued that the process of coming to learn for oneself what makes one better off has its value, but this value can be outweighed by other things, such as the consequences of the mistakes made along the way and the benefits of achieving what makes one better off via a much shorter route. Second, it has been disputed that "structuring choice environments *prevents* people from trying out different" choices than the choice architecture favours (Blumenthal-Barby, 2010, p. 14). Third, under a different definition of rationality, one that focusesless on autonomous reasoning and more on attaining one's ends, nudges can be said to systematically help rather than hurt autonomous decision-making since they help attaining one's ends (Schmidt, 2019, p. 536). Fourth, Schmidt asks: "to form a steely will, should we be exposed to as much temptation as

possible?" (p. 536). His point is directed against the argument of learning. He maintains that for agents to learn their preferences and train their willpower, it is unnecessary for them to live in a world full of temptation and devoid of helpful choice architecture as people can also develop their preferences in a more helpful world. Schmidt therefore argues that infantilization is not a genuine worry. Fifth, Heilmann (2014) notes that a genuine opt-out could mitigate worries about infantilization. He argues that if people are free to opt-out of nudges, they are also free to make their own choices. As long as people can maintain this freedom of choice, they will not become infantilized.

2.6.2 Mistrust

The issue of mistrust states that nudges can cause people to mistrust choice architects. Avitzour et al. (2019) raise the issue that when physicians nudge their patients, this could decrease patient trust in their physicians. They surveyed both doctors and patients, presenting them with scenarios involving health nudges. They found that patients reported significantly more harm to the trusting relationship with their physician than doctors did. Ploug and Holm (2015) add that especially over the long-term patients may come to distrust the entire healthcare system if they feel their preferences are systematically neglected and overridden. In the policy domain, Button (2018) similarly upholds that a nudging government may expect "reactance" and risk losing the trust of the public. Likewise, Ewert (2019a) argues trust is vital for a functioning state-citizen relationship and nudges may jeopardize this relationship when citizens feel they are nudged towards ends they reject. Clavien (2018) agrees governmental nudges may foster mistrust and adds that a nudging government may undermine its own communications. However, Button (2018) stresses that more empirical research is needed to establish such an effect.

2.7 Ethical issue: Democracy and deliberation

The issue of democracy and deliberation is concerned with the worry that nudges weaken democratic institutions. Governmental nudges, their justificatory strategy libertarian paternalism, and more broadly the idea of using behaviourally inspired policies as envisaged in the field of Behavioural Public Policy, have been criticized for hurting democracy and the fundamentals on which it rests such as public deliberation, civic capacity, and political legitimacy (Ewert, 2019b). In this section, we discuss four worries of democracy and deliberation, namely political transparency, deliberation, the slippery slope argument, and excess responsibility. While these worries could be viewed as a form of long-term adverse effects, we will treat them here as separate for three reasons. First, these arguments are exclusively relevant for the government and

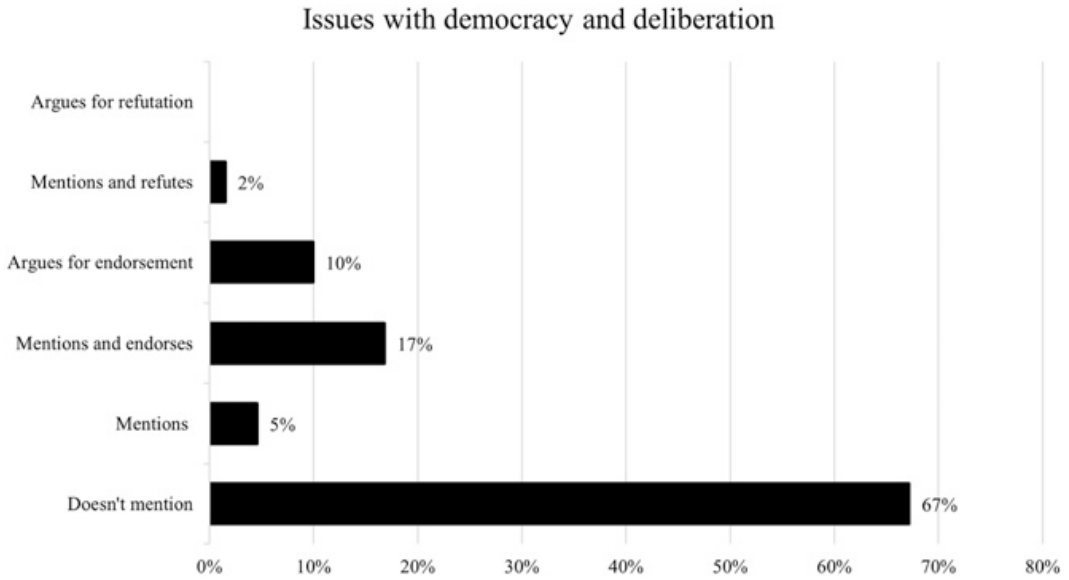


Figure 9 - Overview of how records stand towards the proposition that nudges harm democracy and deliberation.

its treatment of citizens. Second, these arguments can only be understood in collective terms. Third and most importantly, taking the issues of democracy and deliberation separately does justice to the literature in which political problems are also treated rather separately.

2.7.1 Political transparency

The issue of political transparency states that it is wrong for governments to implement nudges without enabling the public to scrutinize those nudges. In the section on manipulation, we have briefly touched on the importance of transparency. Here, we will focus on its political relevance, which is linked to other political issues because it is a prerequisite for "accountability", "respect", "deliberation", "consent", and "acceptance" (Hansen & Jespersen, 2013, p. 15).

For political legitimacy, it is fundamental the public must observe, understand, and deliberate all policies and laws (Farrell, 2015). Only in this way the government can be made accountable, providing the necessary legitimacy for its actions (Yeung, 2016). However, nudges jeopardize democratic control because they are opaque and can be implemented without proper support, creating a situation in which citizens are unable to challenge the government. In essence, the issue of political transparency thus states that nudges escape democratic legitimacy due to their lack of transparency. There are various forms of transparency. The notion of transparency was put forward by Thaler and Sunstein (2008), who proclaimed all nudges should adhere to Rawls' 'publicity principle', which they understand as a requirement for governments to be "able and willing" to publicly defend each nudge they deploy (Thaler & Sunstein, 2008, p. 244). Bovens (2009) discusses what type of transparency would be required for democratic legitimacy. He identifies two stronger notions of transparency by distinguishing between type- and token transparency. Type transparency requires the government to be transparent about the fact that it uses nudge techniques. Token transparency is more demanding since it requires every instance of a nudge can be detected. Bovens contends that there should be "*in principle* token transparency", which means that every single instance of a nudge is discoverable for "a watchful person". In this way, watchful people would be able to inform others if deemed necessary, making nudgers accountable to the public.

Several authors agree that in principle token transparency is sufficient for democratic control (Grüne-Yanoff, 2012) as the essential reason to provide transparency is that potential abuses of its mandate can be observed (Schmidt, 2019). In principle token transparency provides this warranty. However, others propose more far-reaching measures, such as a mandatory public registry of all nudges (Lepenies & Malecka, 2015, p. 435).

2.7.2 Deliberation

Nudges have been criticized as undemocratic because they are not accompanied by sufficient deliberation or public discussion. It is argued that nudges lack "political legitimacy" and that there is no public procedure for their evaluation (Hukkinen, 2016, p. 7). Policy makers can independently decide to implement nudges "without putting effort into cumbersome legislative and deliberative processes" as they "do not require political support, parliamentary procedure and debate" (Lepenies & Malecka, 2015, p. 430). This amounts to a loss of control for citizens (Gumbert, 2019). Critics stress that the implementation of policies regarding a conception of welfare require a wide conversation in which all parts of society participate in an open discussion (Button, 2018). Citizens must engage in discussions about what brings them welfare, not leave it to experts (Ewert, 2019a; Mols et al., 2015). In a similar vein, Brunon-Ernst (2018) argues that there is a risk that individuals who refuse to adopt widely accepted social norms will be nudged to comply. He argues that for social norms to be legitimately imposed on all, transparency and deliberation are necessary requirements. An additional advantage of deliberation is that it accommodates concerns of constrained epistemic access (Schiavone et al., 2014).

Another critique put forward is that nudges produce adverse effects on public discourse (Martin, 2015). Public deliberation is defined by Martin as the capacity of "citizens in a democracy to know and understand the moral and political problems that define their common lives" (Martin, 2015, p. 452). Martin,

Sawicki (2016), Button (2018) and Ewert (2019a) all argue that nudges frustrate this 'civic' capacity, as nudges focus on irrational persuasion, diverting citizens from deliberation, ultimately undermining public discourse. Gumbert (2019) applies this critique to food policies by arguing that citizens should participate in democratic deliberation about food policies. Nudges prevent this, as they nudge citizens to behave in line with governmental policies rather than inviting citizens to reflect on them. Similarly, Farrell (2015) applies this argument in the debate on organ donation defaults and argues that nudges must be "assessed by reference to the extent to which they promote deliberative participation" (p. 278).

Several authors maintain that nudges can be democratic provided they are subject to democratic control. Schmidt (2019) states that deliberation is important but maintains that nudges can be subjected to public deliberation. Häußermann adds that for nudges to be legitimate, citizens must be "actively involved" in the process of bringing them about, which means that citizens must "actively deliberate and participate in the decision about its aim and purpose, its effects and consequences, and its form and substance" (Häußermann, 2020, p. 57). Similarly, Lindhout and Reniers (2017) mention nudges warrant more democratic control than Rawls' publicity principle. They propose nudges should adhere to Habermas' "democratic communication principles", which for them means that nudges must be "understandable, truthful, sincere and respectful" (p. 254).

2.7.3 Slippery slope

Authors who argue for the slippery slope argument state that nudges which are part of a wider libertarian paternalistic policy lead to a dynamic in which progressively more stringent measures are taken, which ultimately culminates in coercion (Rizzo & Whitman, 2009b). The critique is that while the appeal of libertarian paternalist policies is that they respect freedom of choice, over a

longer period they will create the political support for the bans they were designed to prevent. For example, if the government lacks support for introducing a ban on smoking and introducing a ban without support is considered illegitimate, then under the same rationale there is no justification for policies that will create support for such a ban. Nudges aimed at changing social norms that will eventually lead to support for a ban are therefore unjustified (Coggon, 2020). Coggon draws the conclusion that libertarian paternalism is "deficient on its own terms" for this reason, as it does not acquire democratic support legitimately (p. 142). For a defence of nudges against such criticisms, see Nagatsu (2015).

2.7.4 Excess responsibility

The final democratic issue is the issue of excess responsibility which states that in some cases, a nudge "burdens people with more responsibility that they can (reasonably be expected to) manage" (Alfano & Robichaud, 2018, p. 1). Nudges try to induce people to either save more money, live healthier or make greener choices. By devising a policy which aims to alter citizen behaviour, but respects freedom of choice, the government leaves the ultimate choice to act with citizens. In this way, the critique argues, the state makes its citizens responsible for those political issues (Schubert, 2017a). This critique has been connected to Foucault's concept of responsibilization, which states that by focusing on individual behaviour, the government makes the individual responsible for societal problems. Such a line of reasoning is found in Curchin, who argues that libertarian paternalism "emphasises the personal responsibility of the poor for their circumstances" (Curchin, 2017, p. 231). Two additional adverse effects of excess responsibility have been proposed in the literature. First, Gumbert (2019) suggests it leads to a social mechanism in which people are forced into competition with each other to become better, that is more responsible, citizens. Second, Roberts (2018) suggests that some vulnerable people will not respond to nudges. By addressing their individual responsibilities, libertarian paternalism

risks leaving them behind. The issue of excess responsibility may be seen as a critique on transferring responsibilities from the public to the private domain, especially hurting the most vulnerable. Worries about the relationship between nudges and inequality are also mentioned by Allmark and Tod (2014). A related criticism is that the benefits of nudges will not be spread evenly across the population (Ménard, 2010). Likewise, Lehner worries that nudges may be resistible only for a selective portion of the population, allowing it to "escape the costs while benefiting from the gains" (Lehner et al., 2016, p. 175). Lehner quotes Sunstein and Reisch (2013) who found evidence that nudging works best on the uneducated part of the population.

In contrast to the issue of excess responsibility, various authors argue that nudges rather take away too much responsibility from citizens (Bovens, 2009; Selinger & Whyte, 2011; Schubert, 2017b; Baldwin, 2014; Floridi, 2016). This critique echoes the infantilization issue, which states nudges do not let agents decide for themselves and hamper their possibility to take responsibility for their own welfare.

2.8 Ethical theories

Our third aim with this literature review was to investigate which ethical methods are used in the literature on the ethics of nudging. While some argue the issues surrounding autonomy and welfare can be attributed to the classic opposition between the ethical schools of deontology and consequentialism (Iyer, 2017), others note that autonomy can also be interpreted as a necessary condition for welfare and thus can be accounted for by consequentialists and utilitarian liberals (Barton, 2013; Epstein, 2017). So, welfare and autonomy cannot be completely equated with respectively consequentialism and deontology. The question is how do welfare and autonomy relate to one another? While Thaler and Sunstein maintain that libertarian paternalism offers "a real

Third Way" that incorporates both autonomy and welfare (Thaler & Sunstein, 2008, p. 252), transcending the dichotomy between these two values, others argue a "trade-off" between the two goods remains (Barton, 2013, p. 209). For many authors, nudging itself poses a dilemma between autonomy and well-being. As Engelen formulates it: "Assessing the desirability of specific nudges often requires weighing what matters most: generating desirable outcomes or respecting people's decision-making processes." (Engelen, 2019a, p. 206). We list five arguments that address how autonomy and welfare relate to one another when it comes to nudging.

1) Welfare over autonomy. Some authors argue that while autonomy is valuable, there are situations in which welfare should take precedence. Blumenthal-Barby and Burroughs (2012) argue nudges are *prima facie* wrong because they are manipulative, but this manipulation may be justified when especially important values, goals or reasons are at stake. Examples of reasons that are argued to be so important that they warrant a violation of autonomy are charitable giving for the universal poor (Hobbs, 2017, p. 39), using nudges to stimulate paediatric vaccination (Navin, 2017, p. 54), and using nudges to combat the obesity pandemic (Conly, 2013).

2) Autonomy over welfare. The case for autonomy over welfare has been largely discussed in the autonomy section, as most issues with autonomy were raised against the idea of welfare promoting nudges. The autonomy issues are thus intended as an argument to respect autonomy at the expense of welfare. The explicit precedence of autonomy is most pronounced in Scoccia (2019), who argues many people share "disapproval of unauthorized interference with one's choice making by others, even if the interference resulted in better choices or outcomes as judged from one's own point of view." (p. 95).

3) Autonomy as prerequisite for welfare. Sunstein, one of nudging's most fervent proponents, distinguishes thick from thin autonomy. According to Sunstein, under the thick conception autonomy is considered to be an end in

itself. It cannot be overridden by, or reduced to, other goods such as welfare. Alternatively, under the thin conception, autonomy is "an ingredient of welfare" (Sunstein, 2014, p. 124). Sunstein maintains that under the thin conception, autonomy is not valuable in its own right, but rather because it promotes other goods, such as welfare. A loss of autonomy is then only considered harmful when it leads to a reduction of welfare. This view can be attributed to John Stuart Mill (Floridi, 2016; Barton, 2013). Sunstein favours the thin conception and argues "what really does and should matter is welfare" (p. 134). He even goes as far to dismiss autonomy issues altogether: "in many contexts, the objection from autonomy may be a heuristic and that what we really should care about is welfare" (p. 135).

4) Welfare as prerequisite for autonomy. In the healthcare domain, the argument is made that health (welfare) is prerequisite for autonomy. The idea of this argument is that agents who are unhealthy are unable to live the life they want and are thus restricted in their autonomy. For example, Nys and Engelen (2017) argue that health is an essential prerequisite to lead an autonomous life and Engelen (2019b) calls health "a necessary condition for autonomy" (p. 55). Such arguments can be used to justify health nudges, as they portray health nudges as conducive to both welfare and autonomy.

5) Irresolvable tension. Some authors depict the comparison between autonomy and welfare as a cost-benefit problem (Hofmann & Stanak, 2018). Such a cost-benefit analysis should evaluate improvements in welfare against losses of autonomy (Baldwin, 2014, p. 20). The problem with this approach is that while it sounds intuitively appealing, it is difficult to specify how much of the one must be sacrificed for the other. An explicit comparison requires a single overarching measure of goodness, which seems unattainable. This tension has been relatively untouched in the literature on the ethics of nudging. In the following section containing the discussion, we discuss the relationship between welfare and autonomy in slightly more depth.

Other ethical theories

Most authors discuss the ethics of nudging by using a common-sense approach, moral intuitions or an objective good. This becomes clear from the discussion of the ethical issues. The issues we discussed were formulated in terms of autonomy, welfare, long-term adverse effects, and democracy and deliberation. The discussion treated these four goods as self-evident, without tying them to any moral framework or justifying their significance.

Nevertheless, we found several notable exceptions of records in which ethical theories were explicitly discussed. First, republicanism is used by Hamilton (2018), to argue against the use of nudging to obtain informed consent from patients. Hamilton argues that such nudging violated the republican idea of freedom as it is a form of executing arbitrary influence over the nudgee. Second, Häußermann (2020) addresses the critique that nudging is a form of unwanted interference and proposes to justify nudges by using a contractualist approach. In his proposal, citizens would collectively agree on which nudges they deem desirable to address the autonomy issue. Third, Iyer (2017) argues nudges should not solely be judged in reference to autonomy or welfare and that a virtue ethics approach could generate new insights with respect to the desirability of nudges. Likewise, Stanak (2019) uses virtue ethics to overcome ethical issue and to guide ethical nudging. Fourth, Sarkissian (2017), uses Confucianism to address the question of which influences on moral agents are acceptable. From a Confucianist perspective, he argues that manipulating others is inevitable and that the nudges debate should not focus on whether nudges influence, but rather which influences are acceptable. Fifth, Ram-Tiktin (2018) uses Sen's capability approach to argue that health nudges can improve the nudgees' capabilities and can hence improve their freedom. Sixth and finally, Foucault's ideas about freedom, discipline and normalization are used to address the influence nudges can exert on others (see Dorrestijn & Verbeek, 2013; Brunon-Ernst, 2018).

2.9 Discussion

In the discussion we address four points. First, we discuss the definition of nudge and argue for the importance of using a technical definition of nudge. Moreover, we argue for a minor revision of Saghai's (2013) definition to fit the aims of this thesis. Second, we discuss the (moral) relevance of who the choice architect is that nudges and argue for a focus on private nudges. Third, we discuss the premises underlying the welfare arguments in the nudge debate and discuss how the issue of welfare can be extended to technical private nudges. Finally, we discuss how the ethical issue of autonomy is used in the debate, with a focus on how this issue relates to welfare issues.

2.9.1 A revised, technical definition of nudging

As became clear in our discussion of the definition of nudge, there are several conceptions of nudge. We adhere to Saghai's definition of nudging as "A nudges B when A makes it more likely that B will ϕ, primarily by triggering B's shallow cognitive processes, while A's influence preserves B's choice-set and is substantially noncontrolling (i.e., preserves B's freedom of choice)" (Saghai, 2013, p. 491). This definition encapsulates most debates surrounding nudges. First, it conceptualizes nudges as intentional acts by including in the definition of nudge "A nudges B when A …" (Saghai, 2013, p. 491). This articulates the difference between nudges and choice architecture. Second, this definition limits nudges to those interventions that primarily target shallow cognitive processes, respecting nudge's legacy in the behavioural sciences and emphasizing its cognitive mechanism while acknowledging virtually no nudge will *exclusively* target shallow cognitive processes. Third, this definition constitutes a technical conception rather than a libertarian paternalist conception of nudge. Hereby, nudging is defined as a technique for behaviour modification, regardless of its

(normative) goals. The technical definition is broader than the libertarian paternalist definition, which excludes nudges that are not intended to promote the welfare of the nudgee.

However, we propose one modification of Saghai's definition, which is to remove its reference to substantial noncontrol (that is, to remove the phrase "and is substantially noncontrolling (i.e., preserves B's freedom of choice" (Saghai, 2013, p. 491)). Saghai interprets substantial noncontrol as that nudges must be easily resistible for the average person. However, we argue that including the aspect of substantial noncontrol insufficiently accounts for individual differences. What is noncontrolling for one person is controlling for another. Considering digital nudges, which are often personalized as we argue in chapter 3, the average person is of little significance.

Moreover, Saghai refers to behavioural modification techniques that primarily target shallow cognitive processes but are more controlling as behavioural "prods" (2013, p. 492). For pragmatic reasons, alternating between nudge and prod would complicate the analysis unnecessarily. Also, for behavioural modification techniques that primarily target shallow cognitive processes of which it is uncertain how controlling they are, we would lack a term.

We do agree with Saghai that substantial noncontrol is an important aspect for the moral evaluation of nudges. However, we do not agree this aspect should be incorporated within the definition of nudge. For the reasons listed above, we propose to modify the definition of nudging to: A nudges B when A makes it more likely that B will ϕ, primarily by triggering B's shallow cognitive processes, while A's influence preserves B's choice-set. This definition of nudging will incorporate a broader array of behavioural interventions and allows the important moral discussion on easy resistibility to be held within the boundaries of the concept of nudge.

2.9.2 Private and digital nudges

It has been argued that in addition to the ethical issues, it is morally relevant who the agent is that nudges, the choice architect. According to Bovens (2009), "[i]t matters a great deal who is doing the *Nudging*" (p. 12). There are three arguments about choice architects in the literature.

First, it is argued that it is inappropriate for the government to nudge its citizens. According to White (2013) the government displays an "inappropriate attitude" towards its citizens when it nudges them (p. 103). Hausman and Welch (2010) agree with this point and consider governmental nudges "a form of disrespectful social control" (p. 134). These authors hence argue that it is more appropriate for some agents to nudge, than it is for others. For them, the government is held to a higher standard when it comes to changing the behaviour of others. The government is chosen by the people and ought to respect the wishes of the people. Therefore, they argue, the government is not entitled to interfere with citizen choices by means of a nudge.

Second, three arguments were made about how governmental nudges may lead to special issues. Bovens (2009) connects the question of who the choice architect is to self-binding. He maintains that nudging oneself is perfectly acceptable, but questions whether democratic majorities have a sufficient mandate to nudge everyone in a society. White (2013) connects the agency issue to the earlier discussed epistemic concerns. According to him, the government is an especially bad candidate to nudge due to its inability to discover citizen preferences. Glaeser (2005) uses the earlier discussed idea of the endogeneity of errors to argue governmental nudges may be misdirected more often than private ones. He argues that if errors are indeed endogenous, errors will be greater in the public domain, since public servants have smaller incentives to correct their errors than do private parties.

Third, it is argued that people are more vulnerable to governmental nudging than to private nudging because they expect to be manipulated or

persuaded by private companies, while they do not expect this from governments (White, 2013). A similar argument can be found in Baldwin (2014), who adds that the public knows private parties will use temptation in marketing, which they tolerate and anticipate. Hausman and Welch further extend on this point by stating that the enormous amount of private nudging is tolerable only due its limited effectiveness and the fact that private company nudges conflict with each other and thus, to a certain extent, neutralize each other (Hausman & Welch, 2010).

In contrast, Quigley (2014) maintains these authors "display a selective concern" towards governmental nudging (p. 156). Likewise, Thaler and Sunstein state that "[o]n the face of it, it is odd to say that the public architects are always more dangerous than the private ones" (Thaler & Sunstein, 2008, p. 239). Engelen (2019b) notices this difference in acceptance between various nudgers. He sees no particular relevancy in the agency of the nudger yet emphasises that the main point of importance is that there is a "trusting relationship" between the nudger and the nudgee (p. 56).

All these arguments are concerned with the special concerns with governmental nudging. There has been no research into the special concerns that private nudges may cause. Several authors have even dismissed concerns with private nudges, arguing they "cancel one another out" (Hausman & Welch, 2010, p. 131) or are anticipated by consumers (Baldwin, 2014). In our view, these authors fail to appreciate the power that private nudges can exert.

This is surprising, as many companies use nudges to improve sales and encourage consumers to purchase their products. For example, credit card companies have long used loss aversion, framing effects, defaults, and present bias with gigantic success, especially in the US (Thaler, 2015). Most websites use an array of nudges, like defaulting users into newsletter subscription or acceptance of terms and conditions (Smith et al., 2013), and in marketing the use of nudges like social proof and framing is well established (Cialdini, 2007). Such

for-profit nudges form a significant gap in the literature on the ethics of nudging and should be addressed by future research.

Attention for the power of private nudges is especially required in the domain of digital technology. Digital choice architecture can be easily created, tracked, tested, and adapted (Yeung, 2016). By combining the behavioural insights from nudge with big data and AI, digital choice architectures can be designed for highly effective nudging (Susser, 2019). A potent indication of the potential consequences of digital choice architecture is provided by Facebook's election nudge, which gave users the possibility to signal they had voted to their friends. It is estimated that this social proof nudge could increase voter turnout by 0,6%, enough to impact an election (Grassegger, 2018).

Digital technology has enabled the nudging of consumers on a large scale, causing the need to scrutinize the issues with nudging in the digital economy (Calo, 2014; Lavi, 2018). As big tech is becoming the primary platform for all marketing messages, digital private architectures create new issues. Therefore, we agree with Schmidt and Engelen (2020) that for the debate on the ethics of nudging, digital choice architectures are an important candidate for future ethical research. The literature on the ethics of nudging has only begun investigating the consequences of digital nudges on democracy, politics, and the economy in the digital era.

Crucially, the common assumption that private nudges cancel each other out, as Hausman and Welch maintain (2010), is not merely incorrect for digital nudges—it is fundamentally reversed. In direct opposition to their claim, digital nudges actively amplify each other through the economic incentives of the attention economy. Far from neutralizing one another, digital nudges collectively exert pressure on the same psychological vulnerability: our limited attention. Each digital platform, application, and service employs sophisticated choice architectures designed to capture and retain user attention, creating a cumulative effect that manifests in problematic outcomes such as smartphone

addiction, digital overuse, and preference distortion. We discuss these aspects in greater detail in chapter 3.

2.9.3 Welfare

Having established the significant and understudied power of private nudges, particularly in digital contexts, we now turn our attention to the welfare implications of these interventions. If the power of marketers to shape the preferences of consumers becomes too strong, the economy fails in its fundamental purpose: incentivizing companies to tailor their offerings to the genuine preferences of consumers. This concern is particularly acute in digital environments, where the systematic capture of attention through amplifying nudges creates substantial welfare challenges. Framing the debate in this way illustrates the key role the notion of preferences plays in the literature on the ethics of nudging.

We identify four key areas for future research concerning the relationship between private nudging and welfare that will guide the remainder of this dissertation.

First, the justification of nudges relies on a problematic concept of preferences. Nudges are justified by reference to welfare understood as the satisfaction of informed or purified preferences. However, in the results section of the literature review, it was argued that true preferences do not exist and that even if they exist, it would be impossible to overcome the epistemic challenges of gaining access to them. This is problematic because it renders libertarian paternalism a moral theory without direction. Imagine that policy makers aim to adopt nudges to improve welfare but have no preconceived conception of the true preferences of citizens. How would they acquire this knowledge? Before this question is answered, the satisfaction of informed preferences is an impractical goal if our question is how to live, from a first-person perspective. Therefore, a key challenge for future research is to either a) develop a strategy to determine

true preferences, or b) develop a different justification for nudging, one that does not rely on the satisfaction of true preferences.

Second, contrary to Thaler and Sunstein's claim, nudges are not always easily resistible, and they can impose additional burdens. We agree with Blumenthal-Barby (2010) that the assumption of easy-resistibility is overly simplistic. Behavioural economics has demonstrated how choice architecture influences behaviour (Kahneman, 2011; Thaler, 2015), and resisting these influences often requires significant effort through administrative actions, extra cognitive effort, or defying social norms, which may lead to social retribution (Bicchieri, 2023). In commercial digital environments, where sophisticated choice architecture is specifically designed to be difficult to resist, this burden becomes especially concerning as it creates friction between profit maximization and user welfare. Thus, the claim that nudges leave those wishing to resist them unaffected seems highly implausible, particularly in private nudging contexts.

Third, while many argue that nudges improve welfare, there are cases where they could be used for purposes that undermine it. Specifically, private nudges designed to influence consumer behaviour—such as persuading individuals to purchase non-welfare optimizing goods—could lead to economic harm or exploitation. This perspective shifts the focus from errors as inevitable to errors as induced by external actors. The crucial questions then become: Who benefits from inducing errors? What factors make errors more likely to arise? A comprehensive analysis is needed to identify the incentives and capabilities of those who might intentionally manipulate choices and to explore the long-term consequences of such actions. These concerns require further investigation, particularly with respect to welfare and the conception of preferences, which we address in Chapter 4.

Finally, in the results section, we identified two arguments regarding the long-term adverse effects of nudges: infantilization and mistrust. While most authors analyse nudges as one-shot situations, these arguments look at the

possible adverse effects of nudges over longer time frames. In our view, long-term effects of nudges are important to consider because it is reasonable to expect that people will be nudged regularly and pervasively over their lives. For private digital nudges, these long-term concerns are especially pronounced, as the cumulative effect of attention-capturing choice architectures deployed across multiple platforms creates sustained pressure on user behaviour in a singular direction. Digital users are constantly steered towards increased engagement, usage, and consumption. The digital attention economy creates aligned incentives for nearly all platforms to nudge users in ways that may progressively diminish decision-making capabilities and welfare over time.

2.9.4 Autonomy

Since nudges change the behaviour of users, they naturally raise questions of autonomy; are users whose behaviour has been deliberately altered by a third party acting autonomously? Which conditions need to be true to maintain autonomy? We have seen various answers to this question. Thaler and Sunstein (2008) insist that autonomy is preserved as long as users are made better off "as judged by themselves" (p. 5). They hence take a nudgee's endorsement as sufficient for respecting autonomy. Saghai (2013) takes a different stance, arguing that the nudges must be "noncontrolling" (p. 488). His account stresses the strength of the psychological mechanism used to induce behaviour change, which cannot be such that users have difficulty resisting it.

Another autonomy argument in the literature is that nudging is a form of impermissible manipulation because it influences behaviour in the same way that one would manipulate a tool (Noggle, 2018). It could be argued that nudgers are not interested in deliberating with a nudgee by means of reasons but try to pull "causal levers" (Mills, 1995, p. 101). It could be argued that this is incompatible with respect for people as moral beings.

Bovens (2009) investigates if nudges lead to genuine preference change. He uses Frankfurt's concept of a hierarchy of preferences, and rightfully takes nudges to be directed only at first-order preferences, or behaviour. Nudges may be successful in altering behaviour, without altering second-order preferences. Bovens (2009) uses the example of nudging people to save more for their pensions. When people have no preference to save but are successfully nudged to do so, this may result in a "fragmented self" (p. 8); there is a conflict within their overall preference structure, where lower-order preferences are unaligned with the overall hierarchy as a result of the nudge.

Where Frankfurt emphasizes the importance of reflectively endorsing preferences, and hence alignment between first- and second-order preferences, Bovens emphasises, in line with Jon Elster (2016), that there are additional requirements for autonomy. To illustrate this point, he discusses multiple cases. A preference change that is the result of a nudge may be genuine when someone acts as a result of the nudge (first-order preference change) and then comes to appreciate the newly acquired preference through "habituat[ion] in Aristotelian style", the preference change is genuine and autonomous. In the example of saving for a pension, someone may come to appreciate frugality and have acquired a genuine, authentic new preference, or even value.

However, when agents become dependent on the nudge to save, and never fully internalize the preference, even though they reflectively endorse it, Bovens argues, the preference change is "not resilient" (p. 9) and fails to constitute a genuine preference change. This argument mirrors Jon Elster's requirement for autonomous preferences, which states that preferences must be the result of a conscious process of character planning and cannot depend heavily on external circumstances. For Elster, external circumstances include both the available set of options (which is excluded from the definition of nudge) as well as any form or presentation, or manipulation (which we could define as the choice architecture here). Hence, both Bovens and Elster would disagree with Thaler

and Sunstein that an agent is autonomous when they reflectively endorse their actions after being nudged. In their view, an agent must reflectively endorse the action, also in the absence of a nudge.

This brief discussion of Bovens' view on autonomy and preference shaping serves to illustrate that the relation between preferences, welfare and autonomy is complex and entangled. Especially in any moral theory that stresses welfare and (any form of) desire fulfilment, the way in which preferences come about becomes vitally important. If preferences are not formed autonomously, it can be asked what the value of preference satisfaction is. In chapter 4, we use Frankfurt's concept of preference hierarchy and wholeheartedness, Elster's ideas on autonomous preferences, and J.S. Mill's notion of character building (preference formation) to analyse the relationship between welfare, preferences, and autonomy in greater depth to ultimately arrive at a theory of how to live one's life when confronted with digital nudges.

2.10 Conclusion

Having reviewed the literature on the ethics of nudging, we identified 1693 potentially interesting articles, of which after a selection process 131 we identified as relevant for the literature on the ethics of nudging. After analysing these 131 records, we found that there are four major ethical issues with nudging in the literature: the violation of autonomy, questions whether nudges can in fact improve welfare, long-term adverse effects, and harm to democracy and deliberation. In addition, we analysed definitions and moral theories in the literature on the ethics of nudging.

After this analysis, we argued for a technical definition of nudge to incorporate for-profit nudges in the discussion. We have also argued that Saghai's definition is the most useful definition of nudge, but that this definition must be broadened to include nudges which are not easily resistible. This revised,

broader definition allows to uncover a number of under-researched issues with nudging as it includes nudges with other goals than to improve welfare and nudges that are hard to resist. We have identified three main areas that deserve further attention in the literature on the ethics of nudging.

First, we have highlighted several issues with the current conception of welfare in the literature. Beyond the conceptual and epistemic challenges related to preferences, we argue that a comprehensive ethics of nudge should not only address the arguments supporting nudges as welfare-improving but also consider their potential to cause harm. Specifically, this includes the difficulty of resisting nudges, the flawed justification based on informed preferences, and the potential for nudges to be exploited for purposes that undermine welfare.

Second, we have argued that long-term adverse effects and especially the induction of errors deserve more attention. An elaborate analysis of nudge needs to account for the endogeneity of errors, whether errors originate as a result of inducement, or from other sources. Third, we have argued that private nudges, particularly those with the aim of persuading consumers to purchase a marketer's goods, are under-researched.

The combination of 1) the lack of attention for negative welfare effects of nudges, 2) the underdevelopment of the argument of long-term effects, and 3) the neglect of other choice architects than the government all coincide. The literature's emphasis on using nudges to improve welfare makes the government a natural candidate to apply nudges, leading to a lack of focus on other choice architects such as corporations. However, turning our attention to for-profit nudging by private parties, potentially welfare hurting nudges become more conceivable. Such a view on nudging also makes the incentives for private firms to induce errors in consumers more conceivable. Hence, a shift of focus towards other choice architects makes the under-researched issues of a loss of welfare and of induced errors more prominent.

This is markedly evident in the digital economy. As Zuboff (2019) forcefully illustrates, the nature of the business model of the most powerful players in the digital economy Alphabet, Microsoft, Apple, and Meta, is such that their revenue depends on their ability to persuade consumers to click on their advertisements. They are thus directly incentivized to alter the behaviour of their users. For this reason, the digital economy is an important domain to research the ethics of nudging in general and the inducement of errors specifically. Another reason why the digital economy is an interesting field of research from the perspective of the ethics of nudging is that digital environments are extremely suitable for nudging. The plasticity of digital environments and the possibility to track behaviour pervasively enable the multivariate testing of nudges, which improves their effectiveness. In addition, the ability to personalize these environments enables the creation of personalized nudges, improving their potency even further. Thus, in our view, a shift of focus to welfare-decreasing nudges, long-term effects, and private parties as choice architects offers a promising avenue for further research. In the digital economy, these issues all converge, making it a promising domain for further moral inquiry. Together, these suggestions have the potential to significantly broaden the literature on the ethics of nudging.

2.11 Bibliography

Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L. F., Komanduri, S., Leon, G., Sadeh, N., Schaub, F., Sleeper, M., Wang, Y., & Wilson, S. (2017). Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys*, 50(3), 1-41. https://doi.org/10.1145/3054926

Aggarwal, A., Davies, J., & Sullivan, R. (2014). "Nudge" in the clinical consultation—An acceptable form of medical paternalism? *BMC Medical Ethics*, 15, Article 31. https://doi.org/10.1186/1472-6939-15-31

Alfano, M., & Robichaud, P. (2018). Nudges and other moral technologies in the context of power: Assigning and accepting responsibility. In *The Palgrave*

handbook of philosophy and public policy (pp. 235-248). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-93907-0_19

Allmark, P., & Tod, A. M. (2014). Can a nudge keep you warm? Using nudges to reduce excess winter deaths: Insight from the Keeping Warm in Later Life Project (KWILLT). *Journal of Public Health*, 36(1), 111-116. https://doi.org/10.1093/pubmed/fdt067

Avitzour, D., Barnea, R., Avitzour, E., Cohen, H., & Nissan-Rozen, I. (2019). Nudging in the clinic: The ethical implications of differences in doctors' and patients' point of view. *Journal of Medical Ethics*, 45(3), 183-189. https://doi.org/10.1136/medethics-2018-104978

Baldwin, R. (2014). From regulation to behaviour change: Giving nudge the third degree. *The Modern Law Review*, 77(6), 831-857. https://doi.org/10.1111/1468-2230.12094

Barton, A. (2013). How tobacco health warnings can foster autonomy. *Public Health Ethics*, 6(2), 207-219. https://doi.org/10.1093/phe/pht011

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

Bentham, J. (1996). *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*. Clarendon Press.

Berlin, I. (2017). Two concepts of liberty. In *The liberty reader* (pp. 33-57). Routledge.

Bicchieri, C. (2023). Norm nudging and twisting preferences. *Behavioural Public Policy*, *7*(4), 914–923. https://doi.org/10.1017/bpp.2023.5

Binder, M. (2014). Should evolutionary economists embrace libertarian paternalism? *Journal of Evolutionary Economics*, 24(3), 515-539. https://doi.org/10.1007/s00191-013-0323-7

Binder, M., & Lades, L. K. (2015). Autonomy-enhancing paternalism. *Kyklos*, 68(1), 3-27. https://doi.org/10.1111/kykl.12071

Blumenthal-Barby, J. S., & Opel, D. J. (2018). Nudge or grudge? Choice architecture and parental decision-making. *Hastings Center Report*, 48(2), 33-39. https://doi.org/10.1002/hast.837

Blumenthal-Barby, J. S. (2010). Choice architecture: A mechanism for improving decisions while preserving liberty? In *Paternalism: Theory and practice* (pp. 178-196). Cambridge University Press. https://doi.org/10.1017/CBO9781139179003.010

Blumenthal-Barby, J. S., & Burroughs, H. (2012). Seeking better health care outcomes: The ethics of using the "nudge." *The American Journal of Bioethics*, 12(2), 1-10. https://doi.org/10.1080/15265161.2011.634481

Bonotti, M. (2015). Food policy, nutritionism, and public justification. *Journal of Social Philosophy*, 46(4), 402-417. https://doi.org/10.1111/josp.12129

Borenstein, J., & Arkin, R. (2016). Robotic nudges: The ethics of engineering a more socially just human being. *Science and Engineering Ethics*, 22(1), 31-46. https://doi.org/10.1007/s11948-015-9636-2

Borenstein, J., & Arkin, R. C. (2017). Nudging for good: Robots and the ethical appropriateness of nurturing empathy and charitable behavior. *AI & Society*, 32(4), 499-507. https://doi.org/10.1007/s00146-016-0684-1

Bovens, L. (2009). The ethics of nudge. In *Preference change: Approaches from philosophy, economics and psychology* (pp. 207-219). Dordrecht: Springer Netherlands.

Brunon-Ernst, A. (2018). Use of social norms in the production of legal norms: A genealogical and critical approach to nudges. *Cahiers De Droit*, 59(1), 117-142. https://doi.org/10.7202/1043687ar

Burgess, A. (2012). "Nudging" healthy lifestyles: The UK experiments with the behavioural alternative to regulation and the market. *European Journal of Risk Regulation*, 3(1), 3-16.

Burr, C., Cristianini, N., & Ladyman, J. (2018). An analysis of the interaction between intelligent software agents and human users. *Minds and Machines*, 28(4), 735-774. https://doi.org/10.1007/s11023-018-9479-0

Button, M. E. (2018). Bounded rationality without bounded democracy: Nudges, democratic citizenship, and pathways for building civic capacity. *Perspectives on Politics*, 16(4), 1034-1052.

Calo, R. (2014). Code, nudge, or notice. *Iowa Law Review*, 99, 773-802.

Chandler, J. A., & Gruben, V. (2016). The ethics of persuasion: Evaluating the ethical limits on attempting to persuade families to donate the organs of deceased family members. In R. A. Greenberg, A. M. Goldberg, & D. Rodriguez-Arias (Eds.), *Ethical issues in pediatric organ transplantation* (pp. 63-82). Springer. https://doi.org/10.1007/978-3-319-29185-7_4

Cialdini, R. B., (2007). Influence: The psychology of persuasion. Collins.

Clavien, C. (2018). Ethics of nudges: A general framework with a focus on shared preference justifications. *Journal of Moral Education*, 47(3), 366-382. https://doi.org/10.1080/03057240.2017.1408577

Coggon, J. (2020). Smoke free? Public health policy, coercive paternalism, and the ethics of long-game regulation. *Journal of Law and Society*, 47(1), 121-148. https://doi.org/10.1111/jols.12213

Cohen, S. (2013). Nudging and informed consent. *The American Journal of Bioethics*, 13(6), 3-11. https://doi.org/10.1080/15265161.2013.781704

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

Conly, S. (2013). Coercive paternalism in health care: against freedom of choice. *Public Health Ethics*, *6*(3), 241-245.

Curchin, K. (2017). Using behavioural insights to argue for a stronger social safety net: Beyond libertarian paternalism. *Journal of Social Policy*, 46(2), 231-249. https://doi.org/10.1017/S0047279416000672

Damhus, C. S., Petersen, G. B., Ploug, T., & Brodersen, J. (2018). Informed or misinformed choice? Framing effects in a national information pamphlet on colorectal cancer screening. *Health, Risk & Society*, 20(5-6), 241-258. https://doi.org/10.1080/13698575.2018.1499877

Dold, M. F., & Schubert, C. (2018). Toward a behavioral foundation of normative economics. *Review of Behavioral Economics*, 5(3-4), 221-241. https://doi.org/10.1561/105.00000097

Dorrestijn, S., & Verbeek, P. P. (2013). Technology, wellbeing, and freedom: The legacy of utopian design. *International Journal of Design*, 7(3), 45-56.

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

Engelen, B. (2019b). Ethical criteria for health-promoting nudges: A case-by-case analysis. *The American Journal of Bioethics*, 19(5), 48-59. https://doi.org/10.1080/15265161.2019.1588411

Engelen, B., & Nys, T. (2020). Nudging and autonomy: Analyzing and alleviating the worries. *Review of Philosophy and Psychology*, 11(1), 137-156. https://doi.org/10.1007/s13164-019-00450-z

Epstein, W. N. (2017). Nudging patient decision-making. *Washington Law Review*, 92(3), 1255-1315.

Ewert, B. (2017). Promoting health in schools: Theoretical reflections on the settings approach versus nudge tactics. *Social Theory and Health*, 15(4), 430-447. https://doi.org/10.1057/s41285-017-0036-3

Ewert, B. (2019a). From entitled citizens to nudged consumers? Re-examining the hallmarks of health citizenship in the light of the behavioural turn. *Public Policy and Administration*, 34(3), 382-402. https://doi.org/10.1177/0952076718774612

Ewert, B. (2019b). Moving beyond the obsession with nudging individual behaviour: Towards a broader understanding of behavioural public policy. *Public Policy and Administration*. https://doi.org/10.1177/0952076719889090

Farrell, A. M. (2015). Addressing organ shortage: Are nudges the way forward? *Law, Innovation and Technology*, 7(2), 253-282. https://doi.org/10.1080/17579961.2015.1106105

Felsen, G., & Reiner, P. B. (2015). What can neuroscience contribute to the debate over nudging? *Review of Philosophy and Psychology*, 6(3), 469-479. https://doi.org/10.1007/s13164-015-0240-9

Floridi, L. (2016). Tolerant paternalism: Pro-ethical design as a resolution of the dilemma of toleration. *Science and Engineering Ethics*, 22(6), 1669-1688. https://doi.org/10.1007/s11948-015-9733-2

Fumagalli, R. (2016). Decision sciences and the new case for paternalism: Three welfare-related justificatory challenges. *Social Choice and Welfare*, 47(2), 459-480. https://doi.org/10.1007/s00355-016-0972-1

Gelfand, S. D. (2016). The meta-nudge—A response to the claim that the use of nudges during the informed consent process is unavoidable. *Bioethics*, 30(8), 601-608. https://doi.org/10.1111/bioe.12266

Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology*, 6(3), 361-383. https://doi.org/10.1007/s13164-015-0248-1

Giubilini, A., Caviola, L., Maslen, H., Douglas, T., Nussberger, A. M., Faber, N., Vanderslott, S., Loving, S., Harrison, M., & Savulescu, J. (2019). Nudging immunity: The case for vaccinating children in school and day care by default. *HEC Forum*, 31(4), 325-344. https://doi.org/10.1007/s10730-019-09383-7

Glaeser, E. L. (2005). Paternalism and psychology (Working Paper No. 11789). National Bureau of Economic Research.

Glod, W. (2015). How nudges often fail to treat people according to their own preferences. *Social Theory and Practice*, 41(4), 599-617. https://doi.org/10.5840/soctheorpract201541433

Gorin, M., Joffe, S., Dickert, N., & Halpern, S. (2017). Justifying clinical nudges. *Hastings Center Report*, 47(2), 32-38. https://doi.org/10.1002/hast.688

Grassegger, H. (2018, April 15). *Facebook says its voter button is good for turnout – but should the tech giant be nudging us at all? The Guardian*. https://www.theguardian.com/technology/2018/apr/15/facebook-says-it-voter-button-is-good-for-turn-but-should-the-tech-giant-be-nudging-us-at-all

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. (2018). Boosts vs. nudges from a welfarist perspective. *Revue d'Economie Politique*, 128(2), 209-224.

Gumbert, T. (2019). Anti-democratic tenets? Behavioural-economic imaginaries of a future food system. *Politics and Governance*, 7(4), 94-104. https://doi.org/10.17645/pag.v7i4.2216

Halpern, D. (2015). *Inside the nudge unit: How small changes can make a big difference*. Random House.

Hamilton, P. (2018). A Republican argument against nudging and informed consent. *HEC Forum*, 30(3), 267-282. https://doi.org/10.1007/s10730-017-9343-2

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(1), 3-28. https://doi.org/10.1017/S1867299X00002762

Hansen, P. G., Skov, L. R., & Skov, K. L. (2016). Making healthy choices easier: Regulation versus nudging. *Annual Review of Public Health*, 37(1), 237-251. https://doi.org/10.1146/annurev-publhealth-032315-021537

Hausman, D. M. (2018). Efficacious and ethical public paternalism. *Review of Behavioral Economics*, 5(3-4), 261-280. https://doi.org/10.1561/105.00000090

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

Häußermann, J. J. (2020). Nudging and participation: A contractualist approach to behavioural policy. *Philosophy of Management*, 19(1), 45-68. https://doi.org/10.1007/s40926-019-00117-w

Heilmann, C. (2014). Success conditions for nudges: A methodological critique of libertarian paternalism. *European Journal for Philosophy of Science*, 4(1), 75-94. https://doi.org/10.1007/s13194-013-0076-z

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, 12(6), 973-986. https://doi.org/10.1177/1745691617702496

Hilton, D., Treich, N., Lazzara, G., & Tendil, P. (2018). Designing effective nudges that satisfy ethical constraints: The case of environmentally responsible behaviour. *Mind and Society*, 17(1-2), 27-38. https://doi.org/10.1007/s11299-019-00201-8

Hobbs, J. (2017). Nudging charitable giving: The ethics of nudge in international poverty reduction. *Ethics and Global Politics*, 10(1), 37-57. https://doi.org/10.1080/16544951.2017.1312991

Hofmann, B., & Stanak, M. (2018). Nudging in screening: Literature review and ethical guidance. *Patient Education and Counseling*, 101(9), 1561-1569. https://doi.org/10.1016/j.pec.2018.03.021

Houk, T. (2019). On nudging's supposed threat to rational decision-making. *Journal of Medicine and Philosophy*, 44(4), 403-422. https://doi.org/10.1093/jmp/jhz014

Hukkinen, J. I. (2016). Addressing the practical and ethical issues of nudging in environmental policy. *Environmental Values*, 25(3), 329-351. https://doi.org/10.3197/096327116X14598445991501

Ivanković, V., & Engelen, B. (2019). Nudging, transparency, and watchfulness. *Social Theory and Practice*, 45(1), 43-73.

Iyer, K. (2017). Nudging virtue. *Southern California Interdisciplinary Law Journal*, 26(3), 469-492.

Jennings, B., Wertz, F. J., & Morrissey, M. B. (2016). Nudging for health and the predicament of agency: The relational ecology of autonomy and care. *Journal of Theoretical and Philosophical Psychology*, 36(2), 81-99. https://doi.org/10.1037/teo0000041

John, P. (2018). *How far to nudge? Assessing behavioural public policy*. Edward Elgar Publishing.

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

Kanev, D., & Terziev, V. (2017). Behavioral economics: Development, condition and perspectives. In *Proceedings of the 4th International Conference on Education, Social Sciences and Humanities* (pp. 595-606).

Kapsner, A., & Sandfuchs, B. (2015). Nudging as a threat to privacy. *Review of Philosophy and Psychology*, 6(3), 455-468. https://doi.org/10.1007/s13164-015-0261-4

Karlsen, R., & Andersen, A. (2019). Recommendations with a nudge. *Technologies*, 7(2), Article 45. https://doi.org/10.3390/technologies7020045

Kasperbauer, T. J. (2017). The permissibility of nudging for sustainable energy consumption. *Energy Policy*, 111, 52-57. https://doi.org/10.1016/j.enpol.2017.09.015

Kirchgässner, G. (2014). Soft paternalism, merit goods, and normative individualism. *European Journal of Law and Economics*, 43(1), 125-152. https://doi.org/10.1007/s10657-015-9500-5

Korobkin, R. (2009). Libertarian welfarism. *California Law Review*, 97(6), 1651-1685.

Korthals, M. (2015). Which forms of nudging food decisions are ethically acceptable? In *Know your food* (pp. 179-184). Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-813-1_26

Lanzing, M. (2019). "Strongly recommended" revisiting decisional privacy to judge hypernudging in self-tracking technologies. *Philosophy & Technology*, 32(3), 549-568. https://doi.org/10.1007/s13347-018-0316-4

Lavi, M. (2018). Evil nudges. *Vanderbilt Journal of Entertainment and Technology Law*, 21(1), 1-93.

Lecouteux, G. (2015). In search of lost nudges. *Review of Philosophy and Psychology*, 6(3), 397-408. https://doi.org/10.1007/s13164-015-0265-0

Lehner, M., Mont, O., & Heiskanen, E. (2016). Nudging—A promising tool for sustainable consumption behaviour? *Journal of Cleaner Production*, 134, 166-177. https://doi.org/10.1016/j.jclepro.2015.11.086

Lepenies, R., & Małecka, M. (2015). The institutional consequences of nudging—Nudges, politics, and the law. *Review of Philosophy and Psychology*, 6(3), 427-437. https://doi.org/10.1007/s13164-015-0243-6

Lin, Y., Osman, M., & Ashcroft, R. (2017). Nudge: Concept, Effectiveness, and Ethics. *Basic and Applied Social Psychology*, *39*(6), 293–306. https://doi.org/10.1080/01973533.2017.1356304

Lindhout, P., & Reniers, G. (2017). What about nudges in the process industry? Exploring a new safety management tool. *Journal of Loss Prevention in the Process Industries*, 50, 243-256. https://doi.org/10.1016/j.jlp.2017.10.006

Lodge, M., & Wegrich, K. (2016). The rationality paradox of nudge: Rational tools of government in a world of bounded rationality. *Law & Policy*, 38(3), 250-267. https://doi.org/10.1111/lapo.12056

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

Martin, C. (2015). Nudging the public sphere: A Habermasian perspective on public deliberation as an aim of moral education. *Journal of Moral Education*, 44(4), 440-456. https://doi.org/10.1080/03057240.2015.1087392

Mehlman, M. J., Kodish, E., & Berg, J. (2018). Ethical issues in the use of nudges to obtain informed consent for biomedical research. *IRB: Ethics & Human Research*, 40(3), 1-5. https://doi.org/10.1002/eahr.403001

Ménard, J. F. (2010). A "nudge" for public health ethics: Libertarian paternalism as a framework for ethical analysis of public health interventions? *Public Health Ethics*, 3(3), 229-238. https://doi.org/10.1093/phe/phq024

M'hamdi, H. I., Hilhorst, M., Steegers, E. A. P., & de Beaufort, I. (2017). Nudge me, help my baby: On other-regarding nudges. *Journal of Medical Ethics*, 43(10), 702-706. https://doi.org/10.1136/medethics-2016-103656

Mills, C. (2018). The choice architect's trilemma. *Res Publica*, 24(3), 395-414. https://doi.org/10.1007/s11158-017-9363-4

Mols, F., Haslam, S. 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

Mortensen, A. H., Nordhaug, M., & Lohne, V. (2019). Nudging in nursing. *Nursing Ethics*, 26(6), 1601-1610. https://doi.org/10.1177/0969733018779226

Muldoon, R. (2018). Norms, nudges, and autonomy. In *The Palgrave handbook of philosophy and public policy* (pp. 225-233). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-93907-0_18

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

Nahmias, Y., Perez, O., Shlomo, Y., & Stemmer, U. (2019). Privacy preserving social norm nudges. *Michigan Telecommunications and Technology Law Review*, 26(1), 43-91.

Navin, M. C. (2017). The ethics of vaccination nudges in pediatric practice. *HEC Forum*, 29(1), 43-57. https://doi.org/10.1007/s10730-016-9311-2

Noggle, R. (2018). Manipulation, salience, and nudges. *Bioethics*, 32(3), 164-170. https://doi.org/10.1111/bioe.12421

Nys, T. R., & Engelen, B. (2017). Judging nudging: Answering the manipulation objection. *Political Studies*, 65(1), 199-214. https://doi.org/10.1177/0032321716629487

OECD (2020), *Behavioural Insights and Organisations: Fostering Safety Culture*, OECD Publishing, Paris, https://doi.org/10.1787/e6ef217d-en.

Pettit, P. N. (1996). Freedom as antipower. *Ethics*, 106(3), 576-604.

Ploug, T., & Holm, S. (2015). Doctors, patients, and nudging in the clinical context—Four views on nudging and informed consent. *The American Journal of Bioethics*, 15(10), 28-38. https://doi.org/10.1080/15265161.2015.1074303

Ploug, T., Holm, S., & Brodersen, J. (2014). Scientific second-order "nudging" or lobbying by interest groups: The battle over abdominal aortic aneurysm screening programmes. *Medicine, Health Care and Philosophy*, 17(4), 641-650. https://doi.org/10.1007/s11019-014-9566-9

Quigley, M. (2014). Are health nudges coercive? *Monash Bioethics Review*, 32(1-2), 141-158. https://doi.org/10.1007/s40592-014-0008-5

Ram-Tiktin, E. (2018). The tragedy of the commons and population health: The state's intervention in an individual's actions and choices from a capability perspective. *Journal of Human Development and Capabilities*, 19(4), 438-455. https://doi.org/10.1080/19452829.2018.1471672

Ranchordás, S. (2019). Nudging citizens through technology in smart cities. *International Review of Law, Computers & Technology*, 33(3), 306-329. https://doi.org/10.1080/13600869.2019.1590928

Rebonato, R. (2012). *Taking liberties: A critical examination of libertarian paternalism*. Palgrave Macmillan.

Rebonato, R. (2014). A critical assessment of libertarian paternalism. *Journal of Consumer Policy*, 37(3), 357-396. https://doi.org/10.1007/s10603-014-9265-1

Renaud, K., & Zimmermann, V. (2018). Ethical guidelines for nudging in information security and privacy. *International Journal of Human-Computer Studies*, 120, 22-35. https://doi.org/10.1016/j.ijhcs.2018.05.011

Riley, E. (2017). The beneficent nudge program and epistemic injustice. *Ethical Theory and Moral Practice*, 20(3), 597-616. https://doi.org/10.1007/s10677-017-9805-2

Rizzo, M. J., & Whitman, D. G. (2009a). The knowledge problem of new paternalism. *Brigham Young University Law Review*, 2009(4), 905-968.

Rizzo, M. J., & Whitman, D. G. (2009b). Little brother is watching you: New paternalism on the slippery slopes. *Arizona Law Review*, 51, 685-739.

Roberts, J. L. (2018). Nudge-proof: Distributive justice and the ethics of nudging. *Michigan Law Review*, 116(6), 1045-1067.

Saghai, Y. (2013). Salvaging the concept of nudge. *Journal of Medical Ethics*, 39(8), 487-493. https://doi.org/10.1136/medethics-2012-100727

Sawicki, N. N. (2016). Ethical limitations on the state's use of arational persuasion. *Law & Policy*, 38(3), 211-233. https://doi.org/10.1111/lapo.12059

Schiavone, G., De Anna, G., Mameli, M., Rebba, V., & Boniolo, G. (2014). Libertarian paternalism and health care policy: A deliberative proposal. *Medicine, Health Care and Philosophy*, 17(1), 103-113

Schmidt, A. T. (2016). Withdrawing versus withholding freedoms: Nudging and the case of tobacco control. *The American Journal of Bioethics*, 16(7), 3-14. https://doi.org/10.1080/15265161.2016.1180442

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), Article e12658. https://doi.org/10.1111/phc3.12658

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

Schubert, C. (2017b). Exploring the (behavioural) political economy of nudging. *Journal of Institutional Economics*, 13(3), 499-522. https://doi.org/10.1017/S1744137416000448

Scoccia, D. (2019). Paternalisms and nudges. *Economics and Philosophy*, 35(1), 79-102. https://doi.org/10.1017/S0266267118000093

Selinger, E., & Whyte, K. (2011). Is there a right way to nudge? The practice and ethics of choice architecture. *Sociology Compass*, 5(10), 923-935. https://doi.org/10.1111/j.1751-9020.2011.00413.x

Sharif, A., & Moorlock, G. (2018). Influencing relatives to respect donor autonomy: Should we nudge families to consent to organ donation? *Bioethics*, 32(3), 155-163. https://doi.org/10.1111/bioe.12420

Sims, A., & Mueller, T. M. (2019). Nudge versus boost: A distinction without a normative difference. *Economics and Philosophy*, 35(2), 195-222. https://doi.org/10.1017/S0266267118000196

Smith, N. C., Goldstein, D. G., & Johnson, E. J. (2013). Choice without awareness: Ethical and policy implications of defaults. *Journal of Public Policy & Marketing*, 32(2), 159-172.

Stanak, M. (2019). Nudging in neonatology: Practical wisdom and accountability for reasonableness. *International Journal of Technology Assessment in Health Care*, 35(1), 5-9. https://doi.org/10.1017/S0266462318003732

Sugden, R. (2008). Why incoherent preferences do not justify paternalism. *Constitutional Political Economy*, 19(3), 226-248.

Sunstein, C. R. (2014). *Why nudge: The politics of libertarian paternalism*. Yale University Press.

Sunstein, C. R. (2015). The ethics of nudging. *Yale Journal on Regulation*, 32(2), 413-450.

Sunstein, C. R. (2016). *The ethics of influence: Government in the age of behavioral science*. Cambridge University Press. https://doi.org/10.1017/CBO9781316493021

Sunstein, C. R., & Reisch, L. (2013). Green by default. *Kyklos*, 66(3), 398-402.

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *American economic review*, *93*(2), 175-179.

Susser, D. (2019). Invisible influence: Artificial intelligence and the ethics of adaptive choice architectures. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 403-408). https://doi.org/10.1145/3306618.3314286

Thaler, R. H. (2015). *Misbehaving: The making of behavioural economics*. Penguin Books.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Thorndike, A. N., Riis, J., Sonnenberg, L. M., & Levy, D. E. (2014). Traffic-light labels and choice architecture: Promoting healthy food choices. *American Journal of Preventive Medicine*, 46(2), 143-149. https://doi.org/10.1016/j.amepre.2013.10.002

Vallgårda, S. (2012). Nudge—A new and better way to improve health? *Health Policy*, 104(2), 200-203. https://doi.org/10.1016/j.healthpol.2011.10.013

Viale, R. (2018). The normative and descriptive weaknesses of behavioral economics-informed nudge: Depowered paternalism and unjustified libertarianism. *Mind and Society*, 17(1-2), 53-69. https://doi.org/10.1007/s11299-019-00207-2

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

White, M. D. (2013). *The manipulation of choice: Ethics and libertarian paternalism*. Palgrave Macmillan. https://doi.org/10.1057/9781137313577

White, M. D. (2017). Nudging debt: On the ethics of behavioral paternalism in personal finance. *Journal of Financial Counseling and Planning*, 28(2), 225-234. https://doi.org/10.1891/1052-3073.28.2.225

White, M. D. (2019). Nudging merit goods: Conceptual, normative, and practical connections. *Forum for Social Economics*, 48(3), 248-263. https://doi.org/10.1080/07360932.2016.1196594

Whyte, K. P., Selinger, E., Caplan, A. L., & Sadowski, J. (2012). Nudge, nudge or shove, shove—The right way for nudges to increase the supply of donated cadaver organs. *The American Journal of Bioethics*, 12(2), 32-39. https://doi.org/10.1080/15265161.2011.634484

Wilkinson, T. M. (2013). Nudging and manipulation. *Political Studies*, 61(2), 341-355. https://doi.org/10.1111/j.1467-9248.2012.00974.x

Wolcott, G. (2019). Restricting choices: Decision making, the market society, and the forgotten entrepreneur. *Journal of Business Ethics*, 156(2), 293-314. https://doi.org/10.1007/s10551-017-3560-0

Wright, J. D., & Ginsburg, D. H. (2012). Behavioral law and economics: Its origins, fatal flaws, and implications for liberty. *Northwestern University Law Review*, 106(3), 1033-1088.

Yeung, K. (2012). Nudge as fudge. *Modern Law Review*, 75(1), 122-148.

Yeung, K. (2016). The forms and limits of choice architecture as a tool of government. *Law & Policy*, 38(3), 186-210. https://doi.org/10.1111/lapo.12057

Yeung, K. (2017). 'Hypernudge': Big data as a mode of regulation by design. *Information, Communication & Society*, 20(1), 118-136. https://doi.org/10.1080/1369118X.2016.1186713

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.

Chapter 3 - Digital nudging and the digital nudging imperative of big tech

3.1 Introduction

Nudging occurs in many contexts, but its impact is particularly pronounced in digital environments, where nudges are both pervasive and highly effective. Yet, the ethics of digital nudging remains an underexplored area (Schmidt & Engelen, 2020; Calo, 2014). The discussion has suffered from three hurdles. First, there is a lack of a clear definition of digital nudging. A good definition is important because it prepares meaningful avenues of ethical analysis. Directly related, there are many close synonyms to digital nudging.[4] Second, although multiple research fields have investigated digital nudges or equivalent concepts,[5] they have worked largely independent rather than in cooperation. It is unclear how digital nudging "relate[s] to other research strands in [information sciences] like persuasive technology [...] or [behaviour] change support systems [...]" (Lembcke et al., 2019b, p. 8). Third, ethicists and philosophers have not yet taken up the matter to any significant extent. The ethics of digital nudging has so far mainly been discussed by technology related fields. As a result, the ethics of digital nudges is "a particularly fruitful and underresearched area" (Schmidt & Engelen, 2020, p. 10).

However, researching the ethics of digital nudging is crucial given the growing time people spend online and the sheer number of decisions they make in digital spaces. For example, US teenagers spend approximately seven hours

[4] The synonyms to digital nudging are "hypernudge" (Yeung, 2017, p. 122), "big-data (powered) nudge" (Sætra 2019; p. 3), "big nudge" (Helbing et al., 2017), "online manipulation", "digital manipulation" or "automated manipulation" (Susser et al., 2019a, p. 6), "persuasive (digital) technology "(Fogg, 2003, p. 1), and "persuasive digital advertising" (Williams, 2017, p. 129; Aylsworth, 2020, p. 3).

[5] The fields include law (Calo, 2014; Yeung, 2017, Zarksy, 2019), marketing and business studies (Puaschunder, 2017; Nadler & Mcguigan, 2018), behaviour change support systems (Lembcke et al., 2019a; Lembcke et al., 2019b), persuasive technology (Fogg, 2003), human computer interaction (Mirsch et al., 2017), and science and technology studies (Susser, 2019; Susser 2019b).

daily with screens, excluding schoolwork (Robb, 2019). As people spend more time with digital interfaces, they are presented with more digital nudges. They also disclose more personal information, which optimises digital nudges. So, human behaviour is steered both more often and more powerfully because of digital nudges. Scholars are only beginning to realise the extent to which digital technologies influence human decision-making and thereby society at large.

In this chapter, we aim to facilitate a thorough academic debate about the ethics of digital nudges. To this effect, we first propose a definition of digital nudges consistent with the definition of nudge formulated in the previous chapter. Second, we discuss which special features distinguish digital nudging from other forms of nudging. Third, we analyse how digital nudging is an intricate result of the digital economy. We argue that digital nudges are a necessity in the digital economy, which we refer to as the digital nudging imperative. Finally, we argue that the literature on the ethics of digital nudges and other digital persuasion strategies has been overly focused on the role of designers and governments. Instead, we argue, future research should adopt a more user-centric approach and particularly on education.

3.2 Definition of digital nudging

We defined nudging as follows: A nudges B when A makes it more likely that B will ϕ, primarily by triggering B's shallow cognitive processes, while A's influence preserves B's choice-set. This definition is a slightly modified version of Saghai's definition of nudge (Saghai, 2013, p. 491). Having defined nudging, the question now is what constitutes digital nudging.[6]

The most prominent feature that distinguishes digital technologies from non-digital technologies is their ability to handle information. Digital

[6] We are aware of the definitions provided by Lembkce et al. (2019a; 2019b), Weinmann et al. (2016), Meske & Potthoff (2017), and Meske & Amojo (2020). We do not discuss these definitions here.

technologies are superior to non-digital technologies in the collection, storage, and transport of data because they are fast, efficient, reliable, capable of great complexity, and capable of great scale. Their fast and continuous growth has boosted the amount of information that is collected, used and transmitted (Reinsel et al., 2018). Digital technology is first and foremost information handling technology.

As digital technologies are used to handle information, they shape our information environment by determining which information is available to whom at what time. They do this is two ways. First, digital technologies algorithmically determine how information is collected, stored, handled, and transferred. Second, they have an interface for interacting with the user on which information is presented. How this interface displays information is decided by designers of the digital technology. Every design choice they make influences the user in some way and creates opportunities for digital nudging. For example, visitors of webshops are influenced by the fonts, colours, notifications, testimonials and many more elements of the webshop, as we discuss in a later section.

3.3 Special features of digital nudges

Digital nudges differ from non-digital nudges in several ways. We argue that they can be distinguished by the following five characteristics.

3.3.1 Delivered through digital interfaces

Digital nudges reach users through interfaces on their devices (computer, laptop, tablet, smartphone, smartwatch, etc). As a result, digital nudges are highly visual, interactive, omnipresent.

First, digital interfaces enable nudges that are visually rich and engaging. Empirical research has shown that digital interfaces induce users to engage in

intuitive and fast forms of thinking, the types of thinking that rely on heuristics and are most susceptible to nudges, as a result of their visual nature and amount of information (Benartzi & Lehrer, 2015). As a result, digital nudges are often more potent than their non-digital counterparts.

Second, digital nudges are interactive. Users can click, swipe, drag and modify their interfaces and as such the graphical design is livelier. Making a design interactive or playful can be considered a nudge, as it can motivate users to engage. For example, users looking online for a mortgage may like to drag a toolbar rather than to manually to enter a number when discovering how much they can borrow. Such digital nudges raise user engagement and as such makes digital nudges more appealing and more potent than their non-digital counterparts. In addition, it allows designers to prompt the users to disclose more information.

Third, digital nudges to escape constraints of time and space. Digital nudges can reach users at any moment of the day by means of push notifications, wherever they are, as long as their phone is on. As a result, "firms can increasingly choose when to approach consumers, rather than wait until the consumer has decided to enter a market context" (Calo, 2014, p. 1004). As many users carry their devices almost everywhere, even into the bedroom, digital nudges are ubiquitous, persistent and pervasive.

Conceptualizing digital nudges in terms of their technological delivery aligns with perspectives from Yeung (2017), Susser et al. (2019a, 2019b), and Mirsch et al. (2017), reinforcing the idea that their impact is an inherent part of the architecture of digital systems.

3.3.2 Measurable

Digital technologies enable digital designers to measure how users react to digital nudges. The measurability of digital nudges can be leveraged to obtain data and improve their effectiveness. It is both easy and cheap to measure with

detail how users behave in digital (choice) environments. Even the smallest particulars are measured. For example, on many websites it is measured how long users spend on each page, as well as what they read, view, and listen. Even user mouse movements and keyboard-presses can be tracked. In addition, crucially important to digital marketeers, bigger digital advertisement corporations such as Meta and Alphabet know when users complete a sale. Gmail filters attachments from emails to retrieve information on online sales from digital invoices (Zuboff, 2019). Likewise, all e-commerce websites with Facebook integrations report back to Facebook when a sale has been made. So practically everything users do leaves a digital trace (Solove, 2004).

By analysing these data, inferences can be made about user characteristics. Hence, measurability goes further than measuring the effectiveness of designs and collecting click-through rates. Corporations can use the data they acquired to infer all kinds of user characteristics to the personal level. With advanced machine learning techniques, they can find correlations between data items that would not otherwise be observable and even uncover user characteristics that users have not explicitly shared.

According to the Washington post, Facebook provided at least 98 data points to advertisers who use their platform to accurately target ads to users (Dewey, 2016). These data points include demographic features such as age, gender, language, and relationship status; financial features such as home value, income, and net worth; psychological features such psychological profile, IQ, mood, and time of the day a user is most likely to engage in a certain behaviour like purchasing; life events such as whether the user will have or recently had a marriage, birthday, or child expectancy; and finally user values such as political views, religion, and personal interests. As a result, is has been claimed that “Facebook Knows You Better Than Anyone Else” (Quenqua, 2015; Youyou et al., 2015). So, tech firms can “capture and retain intelligence on the consumer's interaction” with them (Calo, 2014, p. 1003).

Digital designers possess two forms of information. First, information on how effective their designs are, which elements make them effective, and how they compare to other designs. Second, they have information on the user. The combination of these two is then "used to shape the informational choice context in which individual decision-making occurs, with the aim of channelling attention and decision-making in directions preferred by the 'choice architect'" (Yeung, 2017, p. 118).

3.3.3 Adaptable

Digital nudges are adaptable and dynamic. It is much easier and cheaper to change a digital design than it is to change a non-digital design. In addition, the control over digital designs is greater than over non-digital designs, as digital designs do not face similar restrictions in terms of material, cost, and time. This means that digital nudges can be finely tailored and updated to every context. Furthermore, digital nudges can be copied infinitely and reach millions of people at little costs.

As a result, digital designers deliberately construct user interactions with great detail (Calo, 2014). This makes digital designs "the ideal medium for leveraging" the power of measurability (Susser et al., 2019b, p. 7). Every bit of information digital designers acquire can immediately and at scale be leveraged to update the design and make it even more effective at steering user behaviour.

The combination of measurability and adaptability, enables the possibility to split test (or A/B test) design aspects and (the effectiveness of) digital nudges. Split testing involves creating multiple different designs and serving both to statistically equivalent large groups to discover which design is most effective in reaching the goals of the designer. Perhaps the most famous successful example of split testing was Google's endeavour to split test large amounts of marginally different colours of blue for the text colour of the results of their search engine. By finding the optimal shade of blue, Google was able to increase revenue by 200

million dollars (Hern, 2014). This is an example of digital nudging because it preserves the choice-set of users (seemingly, nothing relevant in their choice of links changes) yet, their behaviour is modified. This behavioural modification can only be explained by shallow, system 1, cognitive processes that unconsciously favour certain shades of blue over others. The slight alteration of colours is hence a digital nudge; digital designers make it more likely that users will click a link, primarily by triggering the user's shallow cognitive processes, while their influence preserve the user's choice-set.

Split testing has become standard practice for every larger e-commerce platform. Those platforms typically do not test complete designs but tiny aspects of their designs for effectiveness. As a result, many tests may be performed at the same time. This practice ensures that digital nudges are empirically proven to be effective and can be continuously improved upon.

Digital nudges can even be updated in real-time. Designers can determine the time and place they interact with the user and hence can select the optimal time to digitally nudge. For example, apps can send a push notification when users enter a (specific) store, when they are in the city centre, before they go to bed, when they wake, or when they are in proximity to specific other users. Taking things one step further, Facebook told advertisers it can identify moments when teens are feeling 'insecure' and 'worthless' (Levin, 2017). And it could show specific ads on these moments, like make-up advertisement to girls.

However, leveraging control over time and place in digital nudges need not occur via push notifications. Another technique used in social media apps is to change the order of the timeline based on the time, place, and behaviour of the user. Changing the order of information, while leaving the information itself intact is a classic example of a nudge. Although this may sound mundane, it can have great effect. Designers of social media apps know that users are looking to be stimulated or rewarded in a certain sense. They also know that a scheme in which rewards are variable and unpredictable is more addictive. This form of

variable rewarding is called intermittent reinforcement. Designers leverage this insight by classifying bits of information on user timelines on their estimated impact on the user and by devising optimal rewards schemes. Nir Eyal (2013) notes that timeline-based apps can steer user behaviour by turning the timeline into a slot machine of sorts to make the app more addictive and hence sell more ads.

3.3.4 Personalised

Personalisation is the tailoring of messages, services, or products for individual users based on their behaviour or characteristics. With personalisation, it is possible to present each user with a distinct design in digital environments. This creates the opportunity to design unique digital nudges for each user.

Personalisation is the result of taking measurability to the next level (having knowledge about individual users) and combining it with adaptability (tailoring a design to an individual user based on that knowledge). Personalisation is not exclusive to digital nudges as there are examples of personalised non-digital nudges. For example, one could send out personalised letters with personally tailored nudges. However, personalisation is easier, cheaper and more powerful in digital environments due to the added granularity, scale, and speed.

There are several ways how digital nudging and personalisation can be combined. First, digital designers can personalise the technique with which they nudge users. For example, when a government aims to nudge citizens to vote, they could address one user by leveraging loss aversion and frame a message in terms of not missing out on your vote. And then they could nudge another user by framing a message to leverage authority, such as: the president urges you to vote. In this example, the two citizens have been nudged using different

techniques but with the same goal. Mills calls this type of personalisation "delivery personalisation" (Mills, 2022, p. 153).

Second, digital designers can personalise the objective of their nudges. This technique entails using the same nudge for all users with the aim of generating different results. The most common example would be recommendation engines common on e-commerce websites (Thaler & Tucker, 2013). These engines provide personalised recommendations to users based on big data and AI (Möller et al., 2018). Every user is nudged in the same way (for example with the message: this product was specifically selected for you), yet to purchase different products (Nechushtai & Lewis, 2019). Mills calls this type of personalisation "choice personalisation" (Mills, 2022, p. 153).

Third, Sunstein has suggested that using personalisation can of itself be viewed as a nudge (Sunstein, 2014). For example, including a user's name on a website, like on Netflix, makes the user more loyal the site and can hence be viewed as a nudge.

Fourth, when combining personalisation with adaptability, a new possibility arises, namely, to make websites adaptive to user behaviour in real-time. This technique is called "website morphing" (Hauser et al., 2009, p. 203). Morphing can change the website's appearance, information, or even core functionalities based on the user. To morph the website in a way that is conducive to profit, it must tailor the user's "cognitive style" (Hauser et al., 2009, p. 203), which they define as the user's "preferred way of gathering, processing, and evaluating information" (Hayes & Allinson, 1998, p. 850). Based on the user's previous behaviour, a cognitive style is inferred, and persuasion profile is created. Note that morphing does not require much information about a particular user. It was found that to create an effective persuasion profile, even a few clicks can be sufficient, and the profile can be continuously updated with each addition to the clickstream data.

3.3.5 Invisible

How digital nudges are created, and how the mechanisms behind digital nudging function, is invisible to users. How the information displayed on the interface has been created and selected, remains invisible to users. As users are presented with a mere selection of information, they most likely fail to understand what portion of the total set of information has been made available to them. In addition, users are unaware of how many technologies were used to create the interface, with which technologies those technologies interact, and how many designers (human or machine) have influenced the design. Finally, the user may be unacquainted with the goals of designers.

We distinguish several forms of invisibility. First, technological invisibility occurs because of the connected nature of digital technologies. As digital technologies can be combined and work autonomously, the user is unable to recognise which technologies create the digital experience that is served to her. While the interface is intuitive and easy to use, the workings of technology behind the screen can be invisible (Floridi, 2014). For example, users may not even be aware that their clicks are being recorded, the website they visit is being morphed, or their experienced personalised.

Second, algorithmic invisibility (Mittelstadt et al., 2016) is not created by the combination of multiple technologies – as technological invisibility - but by the complexity of a single technology. Certain algorithms, such as artificial neural networks are infamous for their opacity and incomprehensibility. More problematically, even the engineers of these algorithms might not understand how they function and how they come to decisions.

Finally, users are confronted with informational invisibility. When using a digital technology, users see only a portion of the available information, yet they do not know the size this portion, nor which information remains invisible. In addition, the reasons behind the choices to display certain information in favour of other information are unknown to the user.

Though it has been argued that all nudges can be invisible to the nudgee (Bovens, 2009), digital nudges change faster, can be personalised, and may be based on hidden persuasion profiles. The algorithmic and connected nature of digital technologies obfuscates their workings fundamentally. Therefore, the invisibility of digital nudges is at another level as compared to non-digital ones. Finally, it must be added that the invisibility of digital nudges may be caused by the complexity of digital technologies but may also be deliberately created.

3.4 The digital nudging imperative

A thorough ethical analysis of digital nudges must recognize them not as incidental design choices, but as structural necessities born from the economic logic governing digital markets. As the previous chapter highlighted, much of the existing literature on the ethics of nudging focuses on governmental benevolent nudges, often overlooking the role of private actors and the potential for nudges to cause harm. In the digital economy, however, nudging is not merely an optional feature—it is the inevitable expression of an underlying economic mandate that we term the digital nudging imperative.

This imperative stems from a fundamental transformation in business models based on behavioural modification for commercial gain. As Zuboff (2019) observes, the digital economy has created a new economic logic where user behaviour becomes both the target and product of commercial operations. Within this framework, digital nudges are not merely optional design elements but essential business instruments which are systematically deployed to modify user behaviour in predictable ways. Companies deploy these nudges not primarily to improve user welfare, but to direct behaviour toward business objectives: increased engagement, extended usage time, more frequent interactions, and greater data disclosure. Digital nudging thus emerges as the primary mechanism through which digital markets operate and compete.

3.4.1 The structural forces behind the digital nudging imperative

The Digital Nudging Imperative emerges from three interrelated structural forces that shape competition in digital markets: network effects, economies of scale, and economies of scope. These forces create conditions where firms must continuously optimize their choice architectures through digital nudging or risk market elimination.

3.4.1.1 Network effects

Network effects occur when services grow more valuable as their user base expands. Social networks, marketplaces, and communication platforms all increase in utility as more people join them. This creates powerful first-mover advantages and significant barriers to entry: once Facebook or Amazon establishes dominance, competing platforms face immense challenges. Network effects also generate lock-in, making it costly for users to switch platforms even when they wish to, as their social or commercial connections remain on the dominant network. The result is a winner-take-all dynamic, where capturing and retaining user attention becomes existentially important for platform survival.

3.4.1.2 Economies of scale

Economies of scale in the digital economy operate primarily through data accumulation. Unlike traditional businesses where marginal costs increase with scale, digital platforms experience decreasing marginal costs as they grow—each additional user adds value with minimal additional expense. More significantly, the predictive power of data increases exponentially with scale; certain patterns only become visible in massive datasets. This creates a virtuous cycle for incumbent firms: more users generate more data, which enables more effective digital nudges, which attract more users. As Zuboff notes, this surveillance dividend accrues disproportionately to first movers, creating tremendous

pressure to maximize data extraction through increasingly sophisticated digital nudging.

3.4.1.3 Economies of scope

Economies of scope allow digital firms to leverage data and behavioural insights across multiple markets and services. This explains the relentless expansion of major tech firms into adjacent domains. For example, Alphabet moved from search to email, maps, browsers, operating systems, and autonomous vehicles not merely to diversify, but because each new domain generates complementary data that strengthens its core attention-extraction capabilities. These scope economies manifest in the aggressive acquisition strategies of dominant firms, who frequently purchase potential competitors not primarily for their revenue, but for their users' attention and data.

The combined effect of these forces creates what we call the digital nudging imperative: the structural necessity for companies to continuously deploy and refine digital nudges throughout their interfaces to shape user behaviour. This is not merely a strategic choice but an economic requirement for survival in digital markets. Companies that fail to effectively design their choice architecture to nudge users toward increased engagement face extinction in a marketplace, where capturing and directing attention is essential for profitability. In the digital economy, a company’s revenue is directly related to the time a user spends on a platform and hence to the amount of attention they generate. This is perhaps best exemplified by the remarks of ex-Netflix CEO Reed Hastings, who argued that sleep is their biggest competition (Sulleyman, 2017). Netflix knows they capture more market share if users spend more time on their platform. This is the reason they created the infamous design element where the next episode of a series automatically starts, after an episode ends. This design element rationally shouldn't matter to how many episodes users prefer to watch (their choice-set is unaltered), yet in reality it does affect user behaviour. This

design choice is hence a perfect example of a digital nudge, where the behaviour of users is altered by targeting shallow cognitive processes and by leveraging inertia, while preserving their choice-set.

3.4.2 Manifestations of the digital nudging imperative

We now turn to three key manifestations of the digital nudging imperative: the markets for operating devices, online advertising, and e-commerce.

3.4.2.1 Operating systems and devices

Apple, Microsoft, and Alphabet sell operating systems (iOS, Windows, and Android) that are the foundation of digital devices. These systems determine which software can be installed and how data flows through the device, making them a crucial part of the digital economy. These companies are highly dominant in terms of market share and volume. Android, for example, is the default system on millions of Samsung devices worldwide. Control over an operating system extends beyond technical functionality; it shapes user choices and behaviours through digital nudging in at least four key ways.

First, when setting up a new device, such as a smartphone, users are prompted to make choices regarding how they want the device to function, including important privacy settings that determine which companies gain access to their data. These choices are not presented neutrally, the design favours certain choices over other and hence users are nudged here. This nudging occurs via defaults and coloured indicators of preferred settings. For example, when setting up a new Samsung smartphone, which runs on Android, users are nudged to accept Google's preferred settings.

Second, based on these settings, the operating device directly shares user information such as diagnostics and location with the vendor of device and possibly other companies. For example, on Microsoft devices with the Windows

10 operating system, almost everything the user does is tracked (Kalia, 2016), this includes all input of text and voice, URL's from serving the web, program use, and session duration (Zuboff, 2019). All these data can then be used to digitally nudge users further.

Third, control over the operating system enables control over the devices preferred apps. For example, Microsoft was able to develop Internet Explorer into the default browser due to its dominance in the market of operating systems for personal computers. This led to a major fine in 2013 for violating antitrust laws. Another example is that Android devices come with google chrome, YouTube, and google maps pre-installed, or at least recommended (providing a digital nudge to choose these software applications). These apps can then be used to collect even more data. Most importantly, Alphabet's chrome browser and Microsoft's edge browser can be used to track all URLs of web pages that a user visits, which gives an unprecedented amount of user data to these companies (Zuboff, 2019) which can in turn be used to digitally nudge users.

Fourth, control over the operating system provides control over the app store, which is a huge source of revenue as well as, again, a source of user information. Furthermore, it provides a position of power over the applications installed on the device, which is vitally important. For example, Apple has changed its privacy settings and now every time a user downloads an app through Apple's app store which seeks to monitor and transmit user information, the user is prompted to either give permission for this, or decline. If the user declines, the app is unable to collect any user information. Users are not nudged to prefer any choice. Many users opt to decline. There are estimates that this cost Meta around 10 billion USD in revenue loss. As a result, Apple's design decision to place a simple pop-up question before their users has financially hurt Meta significantly, as Meta's revenue model is dependent on those user data. This example also illustrates how one's control over operating systems impacts other contexts. Most importantly for the digital economy: online advertising.

3.4.2.2 Online Advertising

One of the most important, and most scrutinised, revenue models in the digital economy is that of online advertising. Online advertising has in the span of 20 years grown into a multibillion industry (Economist, 2022). It is a highly diverse and widespread market as ads can be placed adjacent to many forms of content and can take many forms. This market is inhabited by social media platforms and companies such as Facebook (owned by Meta), Instagram (owned by Meta), X (formerly Twitter), LinkedIn (owned by Microsoft), Reddit, Snapchat, Pinterest, and TikTok; search engines like Google (owned by Alphabet) and Bing (owned by Microsoft); video streaming YouTube (owned by Alphabet) and Netflix (which recently announced to enter this market by offering cheaper subscriptions which interrupt videos for ads); Audio platform Spotify, messaging services like Gmail (owned by Alphabet) and WhatsApp (owned by Meta); and many blogs, news outlets and content creators.

Among these companies, Meta and especially Alphabet are the largest. In fact, they are so dominant that they can place analytics tools on many smaller websites, measuring all user interactions with those websites, including sales. This is immensely valuable data for two reasons. First, they can analyse which of their ads convert to purchases, allowing them to improve their ads. Second, when they know an ad campaign has led to a sale, they can stop providing ads for that product, making their ads even more effective.

Meta and Alphabet (and other social media and search engines) are particularly successful in the market for online ads as they combine two key points. First, users spend a lot of time with their services, so there is a plenty of room for ads. Second, users reveal many personal details, thereby allowing ads to be targeted.

The revenue models and competition dynamics in this market must be analysed through the lens of attention. Attention is central to the market of digital advertising in two ways. First, more user attention means there is more

room for ads. For example, on YouTube, each additional video watched means an additional chance to display an add. Every additional minute a user's eyeballs are on the screen, the revenue of the recipient of the information grows. Second, more attention leads to the capture of more user data. With each additional action users perform in an app, they reveal more. These data can in turn be used to target ads more effectively. Yet, it can also be used to learn what captures users' attention best and to make the service even more appealing or addictive. So, for digital advertisement companies, or "the attention merchants" as the title of Wu's book goes (Wu, 2017), attention is a knife that cuts both ways.

In the non-digital economy, suppliers are not directly incentivised to - nor can they properly - measure the amount of time in which consumers spend with their products or how they use them. In the digital economy, this is completely different; revenue directly depends on the amount of time users spend on the digital platforms. Hence, digital nudging is a key part of the online advertisement market. First, companies go to great lengths to capture the attention they seek, and digital nudging is an important part of their arsenal of techniques. According to Sunstein "a fact about the human brain [...] is that we have limited attention. [...] A nudge can get us to pay attention." (Fusaro and Sperling-Magro, 2021, para.2).

An important example of attention-grabbing digital nudge is the re-ordering of information, such as happens in search engine results or posts on a social media timeline. Another important aspect of the digital choice architecture is to remove all friction. Many social media apps and video platforms have minimalistic designs to remove all distractions and keep the user zoned in to consuming the content. Furthermore, social media apps use push notifications to recapture the attention of their users when they are not consuming. Second, as in the context of operating systems, many social media and search engines apps use digital nudges to steer users into accepting their preferred settings. They are greatly incentivised to influence users to accept settings that allow advanced

data accumulation and personalisation. Digital nudges are an effective way of acquiring the needed permission to do so. Third, the online ads themselves use nudging techniques to persuade users to click on them and to buy the advertised products. The digital nudging techniques used in online ads include framing effects, using authority (for example people wearing white physicians coats in toothpaste commercials), invoking scarcity (Booking.com was fined for artificially presented hotels as almost full), as well as using the herding effect and social proof (almost every website has reviews by other users to show social proof). These are all examples of digital nudges as they leave the choice set unaltered but still manage to influence user behaviour by targeting shallow cognitive processes.

3.4.2.3 E-commerce

E-commerce provides an ideal environment for digital nudging. Webstores are designed to shape consumer behaviour through interface elements that nudge users toward certain decisions. Unlike traditional retail environments, where physical constraints and human interactions play a role in decision-making, digital marketplaces are adaptable, allowing continuous optimization of choice architecture. Every aspect of an online store, from its layout to its checkout process, can be refined to maximize engagement and conversions.

Digital nudges in e-commerce are pervasive and require minimal effort to implement. Standard web design templates, that every entrepreneur can use without effort, include features that leverage well-documented cognitive biases to influence purchasing behaviour. Common examples include testimonials that invoke social proof, free content that triggers reciprocity, and sign-up prompts that exploit the foot-in-the-door technique. Strategic use of colour and framing guides user perception, while default settings, such as pre-selected newsletter subscriptions, take advantage of inertia. Price anchoring and the decoy effect shape how users evaluate options, while progress bars encourage commitment

by signalling incomplete tasks. Limited-time offers create urgency through scarcity, and trust signals, such as security badges or endorsements, appeal to authority. These elements are not merely persuasive tools but integral components of the digital shopping experience.

Amazon exemplifies the power of digital nudging in ecommerce by integrating data-driven choice architecture into every aspect of its platform. As a marketplace rather than a traditional vendor, Amazon mediates transactions between buyers and third-party sellers. As a marketplace, Amazon's most important asset is user data. This data enables highly personalized recommendations, dynamic pricing strategies, and interface adjustments that subtly steer user decisions. Through features such as "Customers also bought," "Frequently bought together," and tailored homepage suggestions, Amazon constructs a purchasing environment that continuously reinforces engagement and spending.

A key factor in Amazon's success is its extensive use of A/B testing. The company systematically refines digital nudges by analysing user behaviour in real-time and experimenting with alternative designs. Elements such as call-to-action buttons, checkout flows, and promotional placements are continuously adjusted based on empirical evidence of what maximizes conversions. This iterative process ensures that Amazon's interface remains optimized for influencing consumer choices while further refining its data-driven recommendation systems. By seamlessly integrating digital nudging with large-scale data collection, Amazon exemplifies how e-commerce platforms can digitally nudge user behaviour at an unprecedented scale.

3.4.3 Consequences of the digital nudging imperative

3.4.3.1 Nudge stacking

The discussed dynamics of the digital economy and the ensuing ubiquity of digital nudges lead to a specific ethical problem with digital nudging, namely "nudge stacking" (Coons & Weber, 2013, p. 21). Nudge stacking is defined by Invankovic and Engelen as "the danger of too many nudges suffocating the agent in his attempts to navigate influences" (Ivankovic & Engelen, 2019, p. 63). Due to the dynamic of the digital market described in the previous section, digital nudges particularly are ever present in terms of time, location, and volume. Therefore, digital nudging is especially liable to the risk of nudge stacking.

One single nudge can be more easily resisted than an abundance of nudges. When a nudge stimulates nudgees to engage in some action, nudgees may be tempted to follow along with the nudge, taking the path of least resistance. Halfway through the act, nudgee may come to realise that their actions are not in line with their initial commitments and correct themselves. However, this can more easily be done when nudged occasionally yet may become increasingly tiresome if the number of nudges increases.

With nudge stacking on the other hand, problems arise. Resisting many nudges poses a burden on people. It has been argued that nudges can exert mental, emotional, and social pressures over people. Authors who want to emphasise these forms of cognitive costs have described them as a "psychological tax" (Kirchgässner, 2014, p. 12) and as an "aggregate burden" (Baldwin, 2014, p. 16). It stands to reason that these burdens increase when the number of nudges increases. In fact, the burden imposed by a single nudge may be barely noticeable, while the burden of a sea of nudges may prove troublesome.

Another reason that nudge stacking is worrisome is that nudge stacking can push people gradually away from their core values. If nudges constantly move nudgees towards behaviours that are slightly inconsistent with their core

values, nudgees are likely to update their beliefs and values to rationalise their behaviour to themselves and restore consistency between their actions and their beliefs (Bovens, 2008). This means that a constant stream of nudges can push people towards believes or values that are not fundamentally their own.

This is especially true with digital nudges, as the system behind them ensures digital nudges are all in the same direction. If nudges were random, and people would sometimes be nudged to move left and sometimes to move right, nudges might cancel each other out. However, due to the incentives in the digital economy, all nudges are meant to spend more time with devices, devote more attention to digital designs, and to spend more money online. There are no nudges to counterbalance these influences.

3.4.3.2 Surveillance infrastructure

Surveillance infrastructure underlies the entire system of digital nudging. As Susser et al. (2019b) observe, "pervasive digital surveillance puts our decision-making vulnerabilities on permanent display" (p. 6). The continuous monitoring of user behaviour provides the data necessary for increasingly precise behavioural prediction and influence. Unlike traditional market research, which offers periodic snapshots of consumer preferences, digital surveillance provides continuous, granular insights into psychological vulnerabilities and decision-making patterns. This surveillance infrastructure enables what Calo (2014) describes as market manipulation, where firms systematically exploit cognitive biases to maximize engagement and revenue.

The opacity of this surveillance infrastructure amplifies its effects. Users remain largely unaware of how their behaviour is being monitored, analysed, and influenced, limiting their ability to critically assess the forces acting upon them. Digital technologies are “in a real sense hidden" (Susser et al. ,2019a, p. 34). This concealment further undermines users, as they cannot effectively resist influences which they cannot detect.

3.5 A user-centric approach focused on secondary education

In chapter 1, we discussed Tiberius' first-person view to answer the question of how to live. Our goal is to provide recommendations for doing this when confronted with ethical issues created by digital nudging. To address these ethical problems, we need to distinguish between three broad groups of stakeholders: the creators of digital nudges (digital designers and tech companies); the consumers of digital nudges (users), and governments.

In the literature, the focus has been on the creators and governments. The focus on the creators comes specifically from the professional ethics fields, including the philosophy of technology, design ethics and AI ethics (Brey, 2010; Jobin et al., 2019). The focus on governments comes especially from the academic field of law and political science (Yeung, 2017; Lepenies & Malecka, 2015). In contrast, the role of users has received relatively little attention while in our view, pragmatically, the harms of digital nudges could be addressed by a user-centric approach.

Before we continue, an important distinction must be made explicit. While many digital nudges are reliant on big data for their success, we are here not discussing the ethical problems with data collection as such. Instead, we focus on the ethical problems of - and solutions to - using user data to design and deploy digital nudges.

Given that there are ethical issues with digital nudges and given that digital nudges are usually deliberately created by digital designers, it is understandable that to address these ethical issues there exists a reflex to focus on the role of designers. However, there are several reasons why ethical improvements are not to be expected from designers or tech companies.

First, as mentioned above, many digital nudges are not simply designed by a human but are split tested. Advanced machine learning algorithms determine which design elements are most effective. As a result, humans have smaller influence over digital designs or at least frequently lack the power in their institutional contexts to handpick the designs they find pleasing. Of course, leaders of tech firm could reject unethical designs, but automation and split testing make the costs of doing so abundantly clear, and competitive pressures are high.

Second, tech companies are directly incentivised to capture as much attention as possible, as well as to predict and form user behaviour as effectively as possible. Even stronger, the core of their revenue models depends on how effectively they can digitally nudge users into spending more time using their designs and steering them into using certain apps or clicking on certain links or ads. Therefore, while some companies may implement voluntary measures, structural economic incentives make it unlikely that the industry will lead comprehensive reforms of digital nudge practices without external pressure.

Finally, we previously established that individual digital nudges have limited harmful effects in isolation—rather, ethical problems emerge from the cumulative impact of many digital nudges working together. This creates a collective action problem for digital designers and companies. From any individual designer's perspective, adding one more nudge to an already saturated digital environment contributes negligibly to the overall harm. Meanwhile, removing their nudges would provide little meaningful benefit to users but could significantly disadvantage them competitively. Each actor faces the same dilemma: choosing not to implement engagement-maximizing nudges yields minimal ethical gains while imposing financial costs. Consequently, even well-intentioned designers and companies remain trapped in a system where rational individual choices perpetuate collectively harmful outcomes. This makes

industry-wide self-regulation unlikely without coordinated intervention or external or regulatory pressure.

In addition to the constraints that companies face, we argue the government is unlikely to be the source of moral improvements. Governments have limited power to construct effective legal constraints for digital designs that can be created by anyone, in any country. If a Danish user visits a Chinese website that is hosted on a Swiss server, which government will have jurisdiction? Next, elected officials usually lack the technical know-how to properly understand the problems identified above, let alone be able to produce effective regulatory frameworks to adequately tackle them.

Even more limiting to potential digital nudge legislation is that it is nearly impossible to designate nudges in a practical sense. It has even been argued that nudging is inevitable (Thaler and Sunstein, 2008). In their book, they argue that when cafeteria owners realise that the way food is displayed influences the choices of the consumers, it is impossible for them not to nudge their consumers. Whether they display the food randomly, or consciously, their choices will inevitably impact consumer decisions. The same is true of digital designers: no matter which choices they make, they will always influence user decisions. So, if the government would wish to curb digital nudges, it is unclear what exactly their rules would entail. In our view, it is impossible for the government to create a consistent legal framework that effectively addresses the ethical issues with digital nudges.

Against this backdrop, the largest moral improvements in digital nudging are to be expected from users. Users have both the incentives and the potential to limit the harmful effects of digital nudges on their lives. A user-centric approach should shed light onto the most harmful effects on users, and ways users can protect themselves from digital nudges, or at least mitigate the harmful effects they create. For example, the following questions would be meaningful research avenues: which information are users lacking to make them aware of

the problems? Which information and cognitive tools could assist them in recognizing and resisting digital nudges? Which digital customisations such as adblockers, time-line eradicators and internet blockers are most effective? And finally, which (cognitive) strategies can users implement to stay close to their authentic goals and values in the face of digital nudges.

One might object to a user-centric approach that it makes victims of digital manipulation responsible for being manipulated, which is unfair. Yet, once we accept the unfair load on users is a reality, we might still look for moral improvements initiated by them.

Another possible objection is that this approach requires too much of users and that users are unable to stand up to the forces of the digital economy. While we agree that it would be difficult for an isolated user to shield herself from digital harm, we emphasize that a user-centric approach does not require every single user to invent ways to protect herself. There is an important task for journalists, writers, activists and (moral) philosophers to provide helpful information and tools to the public. Likewise, benevolent software engineers can aide users by producing customisability tools to shield unwanted triggers in digital environments.

A particularly promising avenue for research into mitigating the harmful effects of digital nudges is exploring educational approaches for children and teenagers for several reasons. First, young people are growing up in an environment where digital nudges are pervasive, shaping their habits, preferences, and decision-making processes from an early age. With US teens spending 7 hours a day with their screens, excluding schoolwork (Robb, 2019), they represent the demographic most extensively exposed to these influences. Second, developing educational programs that equip young people with knowledge and cognitive tools to recognize and resist digital manipulation could potentially build digital resilience and strengthen their capacity to identify, evaluate, and counter digital nudges. Third, schools are uniquely positioned to

deliver such education systematically, as they can integrate media literacy, critical thinking, and digital self-control strategies into their curricula. Moreover, educational institutions have the infrastructure to tailor these programs to individual students based on gender, culture, age, developmental level, learning preferences, and diverse cognitive needs, including those of neurodiverse students.

Research into equipping students with these skills early on could help them navigate the digital landscape with greater awareness and agency. Boosts represent an especially promising theoretical framework for this approach (Grüne-Yanoff, 2018; Kozyreva et al., 2019). Our focus on education must be understood as the identification for a promising starting point, we do not suggest that other users could not benefit from recommendations to mitigate the harmful effects of digital nudges.

In later chapters, we develop theoretical recommendations for educational policymakers and teachers that could serve as a foundation for future research into addressing the adverse effects of digital nudges.

3.6 Conclusion

In this chapter, we defined digital nudges and, in line with the previous chapter, focused on private digital nudges. We first identified the special features of digital nudges. They are delivered through digital interfaces, measurable, adaptable, personalised, and invisible. Then, we argued that digital nudges are not an incidental addition to the digital economy but a necessary and intrinsic result of it. The digital economy, characterized by monopolistic competition, operates by collecting user (behavioural) data and using that data to influence behaviour. This becomes clearest in three domains, which are dominated by the largest tech firms in the Western world: operating systems and devices, online advertising including social media, and e-commerce.

The result of digital nudges is that users are subjected to intense and omnipresent digital surveillance and a sea of nudges that all point in the same direction. This direction is heavier, habitual interaction, and increased data sharing. This is achieved not only through personalization and adaptation but also through techniques that exploit users' cognitive vulnerabilities, such as intermittent reinforcement and attention-capturing mechanisms. As a result, digital nudges do not merely influence isolated decisions but shape long-term behavioural patterns, making it increasingly difficult for users to disengage.

Finally, we argued that the ethical issues posed by digital nudges must be addressed through a user-centric approach, with a particular emphasis on formal education for children and teenagers. Tech firms cannot be expected to limit the use of digital nudges, as their business models depend on maximizing user engagement through data collection and behavioural influence. Governments face different challenges, and previous literature has already focused on government intervention. Therefore, we argued for an alternative approach. Education provides a concrete and feasible way to help users better understand and manage the adverse effects of digital nudges.

3.7 Bibliography

Aylsworth, T. (2020). Autonomy and manipulation: Refining the argument against persuasive advertising. *Journal of Business Ethics*, 1–11.

Baldwin, R. (2014). From regulation to behaviour change: Giving nudge the third degree. *The Modern Law Review*, *77*(6), 831–857. https://doi.org/10.1111/1468-2230.12094

Benartzi, S., & Lehrer, J. (2015). The smarter screen: What your business can learn from the way consumers think online. Hachette UK.

Bovens, L. (2009). The ethics of nudge. In *Preference change: Approaches from philosophy, economics and psychology* (pp. 207-219). Dordrecht: Springer Netherlands.

Brey, P. (2010). Philosophy of technology after the empirical turn. *Techné: Research in Philosophy and Technology*, *14*(1), 36–48.

Calo, R. (2014). Digital market manipulation. *George Washington Law Review, 82(4),* 995-1051.

Cohen, J. E. (2017). Law for the platform economy. *U.C. Davis Law Review, 51(1),* 133-204.

Coons, C., & Weber, M. (Eds.). (2013). *Paternalism: Theory and practice*. Cambridge University Press. https://doi.org/10.1017/CBO9781139179003

Cullen, C. (2019, August 30). Over 43% of the internet is consumed by Netflix, Google, Amazon, Facebook, Microsoft, and Apple: Globel internet phenomena spotlight. Sandvine. https://www.sandvine.com/blog/netflix-vs.-google-vs.-amazon-vs.-facebook-vs.-microsoft-vs.-apple-traffic-share-of-internet-brands-global-internet-phenomena-spotlight

Dewey, C. (2016, August 19). *98 personal data points that Facebook uses to target ads to you.* The Washington Post. https://www.washingtonpost.com/news/the-intersect/wp/2016/08/19/98-personal-data-points-that-facebook-uses-to-target-ads-to-you/

Floridi, L. (2014). The 4th revolution: How the infosphere is reshaping human reality. Oxford University Press.

Fogg, B. J. (2003). Persuasive Technology: using computers to change what we think and do. Morgan Kaufmann Publishers Inc.

Fusaro, R. & Sperling-Magro, J. (2021, August 6). *Much anew about 'nudging'*. McKinsey & Company. https://www-mckinsey-com.translate.goog/capabilities/strategy-and-corporate-finance/our-insights/much-anew-about-nudging?_x_tr_sl=en&_x_tr_tl=nl&_x_tr_hl=nl&_x_tr_pto=sc&_x_tr_hist=true

Grüne-Yanoff, T. (2018). Boosts vs. nudges from a welfarist perspective. *Revue D Economie Politique*, *128*(2), 209–224.

Hauser, J. R., Urban, G. L., Liberali, G., & Braun, M. (2009). Website morphing. *Marketing Science*, *28*(2), 202–223. https://doi.org/10.1287/mksc.1080.0459

Hayes, J. and Allinson, C.W. (1998), Cognitive style and the theory and practice of individual and collective learning in organizations. *Human Relations, 31*, 7, 847-71.

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. V., & Zwitter, A. (2019). Will Democracy Survive Big Data and Artificial Intelligence? In D. Helbing (Ed.), *Towards Digital*

Enlightenment: Essays on the Dark and Light Sides of the Digital Revolution (pp. 73–98). Springer International Publishing. https://doi.org/10.1007/978-3-319-90869-4_7

Hern, A. (2014, February 5). *Why Google has 200m reasons to put engineers over designers*. The Guardian. https://www.theguardian.com/technology/2014/feb/05/why-google-engineers-designers

Hills, T. T. (2019). The Dark Side of Information Proliferation. *Perspectives on Psychological Science*, 14(3), 323–330. https://doi.org/10.1177/1745691618803647

Hu, Y.-S. (2019). The impact of increasing returns on knowledge and big data: From Adam Smith and Allyn Young to the age of machine learning and digital platforms. (SSRN Scholarly Paper No. 3414339). https://doi.org/10.2139/ssrn.3414339

Ivanković, V., & Engelen, B. (2019). Nudging, Transparency, and Watchfulness. *Social Theory and Practice*, *45*(1), 43–73. http://www.jstor.org/stable/45218911

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Jolly, J. (2021, February 6). *Is big tech now just too big to stomach?* The Guardian. https://www.theguardian.com/business/2021/feb/06/is-big-tech-now-just-too-big-to-stomach

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kalia; A. (2016, August 17). *With Windows 10, Microsoft blatantly disregards user choice and privacy: A deep dive*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2016/08/windows-10-microsoft-blatantly-disregards-user-choice-and-privacy-deep-dive

Kirchgässner, G. (2014) Soft paternalism, merit goods, and normative individualism. *European Journal of Law and Economics*, 43(1), 125–152. https://doi.org/10.1007/s10657-015-9500-5

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, *21*(3), 103-156.

Lembcke, T.-B., Engelbrecht, N., Brendel, A. B., Herrenkind, B., & Kolbe, L. M. (2019a). Towards a Unified Understanding of Digital Nudging by Addressing its Analog Roots. *PACIS 2019 Proceedings*. https://aisel.aisnet.org/pacis2019/123

Lembcke, T.-B., Engelbrecht, N., Brendel, A. B., & Kolbe, L. (2019b). To nudge or not to nudge: Ethical considerations of digital nudging based on its behavioral economics roots. In *ECIS*.

Lepenies, R., & Malecka, M. (2015). The Institutional Consequences of Nudging—Nudges, Politics, and the Law. *Review of Philosophy and Psychology*, *6*(3), 427–437. https://doi.org/10.1007/s13164-015-0243-6

Levin, S. (2017, May 1). *Facebook told advertisers it can identify teens feeling 'insecure' and 'worthless'*. The Guardian. https://www.theguardian.com/technology/2017/may/01/facebook-advertising-data-insecure-teens

Meske, C., & Potthoff, T. (2017). The DINU-Model – A process model for the design of nudges. In Proceedings of the 25th European Conference on Information Systems (ECIS) (pp. 2587-2597). ISBN 978-0-9915567-0-0 Research-in-Progress Papers. Retrieved from http://aisel.aisnet.org/ecis2017_rip/11

Meske, C., & Amojo, I. (2020). *Ethical Guidelines for the Construction of Digital Nudges*. https://doi.org/10.24251/HICSS.2020.480

Mills, S. (2022). Personalized nudging. *Behavioural Public Policy*, *6*(1), 150-159.

Mirsch, T., Lehrer, C., & Jung, R. (2017). Digital Nudging: Altering User Behavior in Digital Environments. In J.M. Leimeister & W. Brenner (Eds.), Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017), St. Gallen (pp. 634-648).

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1-21. https://doi.org/10.1177/2053951716679679

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, *21*(7), 959-977.

Nadler, A., & McGuigan, L. (2018). An impulse to exploit: The behavioral turn in data-driven marketing. *Critical Studies in Media Communication*, *35*(2), 151–165.

Nechushtai, E., & Lewis, S. C. (2019). What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior*, *90*, 298-307.

Puaschunder, J. M. (2017). The nudging divide in the digital big data era. *International Journal of Research in Business, Economics and Management*, *4*, 11–12.

Quenqua, D. (2015, January 19). *Facebook knows you better than anyone else.* The New York Times. https://www.nytimes.com/2015/01/20/science/facebook-knows-you-better-than-anyone-else.html

Reinsel, D., Gantz, J. & Rydning, J. (2018, November) The Digitization of the World From Edge to Core. IDC & Seagate. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Robb, M. (2019). *Tweens, teens, and phones: What our 2019 research reveals.* Common Sense. https://www.commonsensemedia.org/kids-action/articles/tweens-teens-and-phones-what-our-2019-research-reveals

Saghai, Y. (2013). Salvaging the concept of nudge. *Journal of Medical Ethics*, *39*(8), 487–493. https://doi.org/10.1136/medethics-2012-100727

Schmidt, A. T., & Engelen, B. (2020). The ethics of nudging: An overview. *Journal of Business Ethics*, 162(3), 527-542. https://doi.org/10.1007/s10551-018-4092-1

Solove, D. J. (2004). The digital person: Technology and privacy in the information age (Vol. 1). NyU Press.

Sulleyman, A. (2017, April 19). *Netflix's biggest competition is sleep, says CEO Reed Hastings.* Independent. https://www.independent.co.uk/tech/netflix-downloads-sleep-biggest-competition-video-streaming-ceo-reed-hastings-amazon-prime-sky-go-now-tv-a7690561.html

Sunstein, C. R. (2014). Why nudge: The politics of libertarian paternalism. Yale University Press.

Susser, D. (2019). Invisible influence: Artificial intelligence and the ethics of adaptive choice architectures. *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 403–408. Scopus. https://doi.org/10.1145/3306618.3314286

Susser, D., Roessler, B., & Nissenbaum, H. (2019a). Online manipulation: hidden influences in digital world. *Georgetown Law Technology Review, 4(1),* 1-46. https://doi.org/10.2139/ssrn.3306006

Susser, D., Roessler, B., & Nissenbaum, H. (2019b). Technology, autonomy, and manipulation. *Internet Policy Review*, *8*(2). Scopus. https://doi.org/10.14763/2019.2.1410

Sætra, H. S. (2019). When nudge comes to shove: Liberty and nudging in the era of big data. *Technology in Society*, *59*, 101130. https://doi.org/10.1016/j.techsoc.2019.04.006

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Penguin.

Thaler, R., & Tucker, W. (2013). Smarter Information, Smarter Consumers. *Harvard Business Review*, 91(1–2), 44–54.

The Economist. (2022, September 18). The $300bn Google-Meta advertising duopoly is under attack. *The Economist*. https://www.economist.com/business/2022/09/18/the-300bn-google-meta-advertising-duopoly-is-under-attack

Weinmann, M., Schneider, C., & vom Brocke, J. (2016). Digital Nudging. *Business & Information Systems Engineering*, *58*(6), 433–436. https://doi.org/10.1007/s12599-016-0453-1

Williams, J. (2017). *Freedom and persuasion in the attention economy* (Doctoral dissertation). University of Oxford.

Wu, T. (2017). The attention merchants: The epic scramble to get inside our heads. Vintage.

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, *20*(1), 118–136. https://doi.org/10.1080/1369118X.2016.1186713

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036-1040.

Zarsky, T. Z. (2019). Privacy and Manipulation in the Digital Age. *Theoretical Inquiries in Law*, *20*(1), 157–188. https://doi.org/10.1515/til-2019-0006

Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Profile Books.

Chapter 4 – A process view on preference shaping for the ethics of digital nudging

4.1 Introduction

The behavioural conception of welfare, and the way it is in the literature on the ethics of nudging, are problematic. It is based on refuting the neoclassical economic idea of welfare, which assumes agents are rational decisionmakers who have complete and consistent preferences. For neoclassical welfare economists, welfare is the satisfaction of those preferences. Behavioural welfare instead takes agents to be frequently mistaken, which means they make choices that conflict with their preferences. Proponents of behavioural welfare hold that governments should observe agents and distil their true preferences from their mistaken choices, a process referred to as preference purification. A literature has emerged on mathematical tools to achieve preference purification (Bernheim 2007; Bernheim & Rangel 2007; 2009). Behavioural welfare is then the satisfaction of those purified preferences. Proponents of behavioural welfare argue that governments should nudge agents toward their true preferences to improve their welfare. In addition, they claim agents can reflectively endorse such nudges.

At first sight, purified preferences resemble informed preferences, and behavioural welfare may seem close to informed desire accounts of welfare (see Brandt 1979; Griffin 1986; Railton 1986). However, the claim advanced by behavioural economists is stronger. Informed desire theorists maintain that welfare consists in the satisfaction of the preferences an agent would have if fully informed or ideally rational. They need not assume that such an idealized self is psychologically present in agents, nor that it is attainable. Moreover, they are free to take different stances on how the idealization should function. For example, Railton (1986) proposes that the ideal self should stand in an advisory

relation to the actual self, rather than serve as a direct target for policy (see Smith (2004) for a helpful discussion of the advice model and the example model). Relatedly, informed desire theorists are not committed to the claim that informed preferences are epistemically accessible, either to the agent or to external observers. The question of accessibility is not central to specifying what constitutes the good life.

Behavioural accounts of welfare, by contrast, are explicitly improvement-oriented and policy-facing. Preference purification is intended to guide intervention. In order to justify nudges as welfare-enhancing, policymakers must be able to identify what agents' true preferences are. Behavioural welfare therefore presupposes that purified preferences are, at least in principle, epistemically accessible to third parties. Without such accessibility, the claim that a given intervention tracks an agent's true preferences would lack justification.

This epistemic commitment has normative consequences. If welfare is identified with the satisfaction of purified preferences, and if those preferences are taken to be discoverable through behavioural analysis, then normative authority shifts toward those who possess the relevant epistemic tools. The agent's own reflective perspective risks becoming secondary to externally reconstructed preference orderings.

As argued in the introduction, this thesis is not concerned with formulating a comprehensive theory of the good life, but with offering an account of "how to live [one's] life" from the first-person perspective, in line with Valerie Tiberius' Reflective Wisdom Account (2008, p. 9). On this view, what matters is not the recovery of a hidden, purified set of true preferences underlying behavioural mistakes, but the ongoing activity of reflection through which agents shape and revise their evaluative outlook. Tiberius rejects highly idealized conceptions of the self that detach normative authority from the agent's actual psychological capacities. She rightly argues that a theory that appeals to the

preferences that agents would have under conditions far removed from their lived psychology, risks losing contact with the standpoint from which practical guidance is meaningful.

From the perspective of reflective wisdom, therefore, the appeal to purified preferences appears normatively misdirected. Preference purification attempts to correct behaviour by filtering out cognitive biases and situational distortions in order to approximate what the agent really wants. Yet this model presupposes that beneath observed inconsistency lies a determinate and stable preference ordering waiting to be uncovered. Reflective wisdom, by contrast, treats inconsistency not merely as noise to be mathematically eliminated but as a potential occasion for reflection and value revision. The process by which agents form preferences and make decisions is the relevant point of focus here This process is explicitly held to psychological realities and limitations.

To avoid the problems inherent to idea of behavioural welfare, we argue it should be replaced by a conception of welfare that is based on processes rather than targets or goals. According to the concept of procedural rationality, decisions are deemed rational based on the process used to arrive at them, rather than solely by their outcomes. This means agents can be rational even though they face limitations in information, cognitive abilities, and willpower. Improving upon these three elements leads to better preference shaping, better decisions, and higher welfare.

To illustrate the problems with behavioural welfare, we need to discuss its welfare economic roots. Traditional welfare economics, called neoclassical welfare economics, assumes agents conform to consistency rationality. Consistency rationality means agents have complete and consistent preferences over all available alternatives. It is assumed agent choices perfectly reveal their preferences. Welfare is the satisfaction of those preferences.

Behavioural welfare economics, the academic field that brought about nudging, demonstrates agents seldom conform to consistency rationality. It

argues agents lack the information, cognitive abilities and willpower required for consistency rationality. As a result, agents frequently make mistakes. Their choices do not reveal their true preferences, and they fail to optimize their welfare. Behavioural welfare economics is therefore confronted with the question: if choices no longer reveal preferences, what is welfare? The consensus in behavioural welfare economics is that welfare is the satisfaction of purified preferences. Purified preferences are those preferences that an agent would have revealed if they had unlimited information, cognitive abilities, and willpower.

Purified preferences conform to consistency rationality. Behavioural welfare economics hence discards consistency rationality as an explanatory model but maintains it as a normative standard. Behavioural welfare is the satisfaction of purified preferences, which must conform to consistency rationality. Behavioural welfare hence still relies on consistency rationality.

Behavioural welfare's reliance on consistency rationality as a normative standard creates problems. First, behavioural welfare relies on the assumption that agents have complete and consistent preferences, while it refuses to explicitly endorse this assumption. Second, it conceptualizes preferences as unmoved movers, and it fails to acknowledge preferences need to be shaped. In addition, we identify a problem that stems from behavioural welfare's reliance on preference purification. Behavioural economists who advocate for the use of mathematical tools that should enable the purification of preferences lack an objective method. As a result, purification risks conflating socially desired preferences with true preferences, leading to excessive and harmful paternalism.

As a result of these problems, the dominant welfare conception in the literature on the ethics of nudging must be replaced.

We argue for an alternative conception of welfare based on Herbert A. Simon's concept of procedural rationality. This view acknowledges that agents often lack the information, cognitive abilities, and willpower to reach consistency rationality. Yet, in contrast to behavioural welfare, we argue that

consistency rationality is no relevant benchmark. Agents don't need their preferences to be purified to find out what they truly prefer. Instead, agents must aim to improve their information, cognitive abilities, and willpower so that they can mitigate the harmful welfare effects created by digital nudges.

This chapter is structured in three sections. The first section presents an overview of neoclassical welfare economics since this field is the root of the welfare conception dominant in the literature on the ethics of nudging. In this section we 1) discuss the neoclassical view on rationality, namely consistency rationality; 2) explain the neoclassical view on preferences, namely revealed preferences; 3) outline the neoclassical welfare conception; 4) discuss the policy implications of this welfare conception, including its view on nudges; 5), critically evaluate the neoclassical welfare conception based on consistency rationality. The second section follows a similar outline for behavioural welfare economics. This section also argues that neoclassical and behavioural welfare economics share important features and - as a result - important problems. In the third section, we conceptualize welfare based on Simon's account of procedural rationality (Simon, 2001). Here, we argue, the focus must be on improving information, cognitive abilities, and willpower. These three elements are the basis for evaluating the ethical implications of digital nudging.

4.2 Neoclassical welfare economics

The dominant welfare conception in the literature on the ethics of nudging cannot be understood properly without understanding where it came from: neoclassical welfare economics. Neoclassical welfare economics is founded on consistency rationality.

4.2.1 Consistency rationality

Consistency rationality has two components (Elster, 1991). Normatively it stipulates how agents should behave given their preferences, to achieve them. Positively then, the theory can explain, describe, or predict behaviour by assuming agents follow the normative stipulations (Elster, 1991). Consistency rationality is the foundation of neoclassical economics, which indeed explains behaviour by assuming all agents follow its normative stipulations (Hausman, 2011).

Agents are consistency rational when their preferences conform to four axioms. The first axiom holds that (1) preferences determine choices. This axiom can be referred to as either "choice-determination" (Hausman, 2011, p. 12) or as "utility-maximisation" (Hausman & McPherson, 2009, p. 5). In addition, (2) preferences are "complete" which means that for any two given alternatives x and y, agents must either prefer x over y, y over x, or be indifferent between the two (Hausman & McPherson, 2009, p. 5). (3) preferences are "transitive" which means that for three alternatives, x, y, and z, if agents prefer x over y and y over z, they must prefer x over z (Hausman & McPherson, 2009, p. 5). (4) preferences are "context-independent" which means that "whether agents prefer x over y, remains stable across contexts" (Hausman, 2011, p. 16). While it can vary per the author, context-independence usually means that whether agents prefer x over y, is not affected by irrelevant factors, such as a third alternative, z, becoming (un)available or the description of alternatives.

4.2.2 Revealed preferences and expected utility

Based on these four axioms, Paul Samuelson (1938) created revealed preference theory. This theory in essence provides an empirical test of consistency that checks if choices[7], rather than preferences, satisfy the axioms

[7] We use the terms actions, choices, and behaviour interchangeably in this context.

of rationality. If this is the case, preferences are perfectly revealed by choices, hence the name revealed preference theory. The advantage of revealed preference theory is that it does not need to make assumptions with respect to the unobservable notion of utility and can focus its analyses on observable choices instead. Many economic models are based on revealed preferences, which is foundational to neoclassical economic theory (Read, 2004). Revealed preference theory has substantial normative implications, as it grounds welfare analysis in actual behaviour, asserting that the choices individuals make are indicative of their utility.

So far, we have considered deterministic scenarios; agents formulate preferences over alternative states of the world and choose and obtain their most preferred option. However, many preferences are over alternative future states of the world in which agents face risk; the future they prefer may not be realised. Expected utility theory, first formulated by Von Neumann and Morgenstern (1947), shows how choices involving risks can be described as lotteries or gambles. In situations involving risk, agents determine their choices by comparing the expected utilities of the alternatives, that is they compare "the weighted sums obtained by adding the utility values of outcomes multiplied by their respective probabilities" (Mongin, 1998, p. 171). We refer to preferences over alternatives that take probability into account as preferences over gambles.

The value of expected utility can be illustrated easily with an example in which expected utilities are quantified in terms of money. When an agent can choose between receiving $100, or a lottery ticket that has a 20% chance of winning $1000 and an 80% of winning nothing, the expected utility of the first option is obviously $100, and the expected utility of the second option is 0,2 x 1000 = $200. The second option has the highest expected utility and is rationally preferred if we assume the decisionmaker is neutral to risk. The choice between these two options can be thought of as a choice between two gambles. Agents form preferences over gambles by comparing expected utilities.

4.2.3 Welfare conception

Neoclassical welfare is equated with utility, defined as the satisfaction of preferences. From this, it follows that "the main normative criterion [in neoclassical welfare is] the satisfaction of individuals' preferences" (Sugden, 2018, p. 6). The desirability, appropriateness, or correctness of economic policies and institutions is usually evaluated based on their conduciveness to neoclassical welfare (Hausman, 2018). For example, whether to execute an infrastructural project, adopt a new tax, or reform the digital market, it has been argued that a change to neoclassical welfare is the proper criterion to base the decision on (Hausman, 2018).

4.2.4 Policy implications

Typically, welfare economics is pragmatic in the sense that it aims to provide a form of usefulness in the sense of improving social welfare (Ng, 2004). To this effect, welfare economists are accustomed to making policy recommendations. In fact, according to Sugden, it is customary for welfare economic texts to be written as if directed at a "benevolent autocrat" (Sugden, 2018, p. 19).

Politicians and policy makers hold economists in high regard (Halpern, 2015). Political decisions and public policies are heavily influenced by the concept of welfare that economists use (Atkinson, 2001). For example, economic arguments regarding the efficient functioning of markets and ideas on how cost-benefit analyses of investments must be performed are based on it.

The dominance of neoclassical economics in policy circles means that consumers are typically regarded as rational utility optimizers. Hence, mistaken consumer choices, or consumers who fail to optimize their own welfare, were scarcely noticed in policy circles. Consequently, "in its normative orientation,

conventional law and economics is often strongly anti-paternalistic. The idea of 'consumer sovereignty' plays a large role; citizens, assuming they have reasonable access to relevant information, are thought to be the best judges of what will promote their own welfare" (Jolls et al., 1998, p. 1541). Consumers are trusted to respond appropriately to incentives and information to fulfil their preferences.

As context-dependent preferences do not exist in neoclassical economics, nudging does not exist, as nudging entails altering the context to influence behaviour. From the perspective of neoclassical economics and consistency rationality, nudges do not exist. Agents are assumed to only act based on their preferences, not on contextual characteristics. If agent preferences would supervene on contextual characteristics, their preferences would conflict across contexts, rendering it unclear what the agent prefers overall and violating consistency rationality.

Moreover, the dominance of neoclassical economic theory in policy circles has meant that consumers are understood to require minimal amounts of protection and guidance. In any case, they need no protection from nudges, as these do not exist.

As a result, neoclassical economics provides no conceptual basis for governments to regulate malicious and welfare-degrading nudges, as 1) nudges do not exists and 2) consumers are unable to make decisions that hurt their own welfare. These ideas regarding nudges stem directly from the theoretical commitments of neoclassical welfare economics which lead to various normative commitments and political ideas. Obviously, this creates problems for a sound ethics of nudging.

4.2.5 Critical evaluation

There are important problems with neoclassical welfare stemming from its reliance on consistency rationality as a normative standard. Later, it becomes

clear these problems are incorporated into behavioural welfare, which is dominant in the literature on the ethics of nudging.

Excess of cognitive demands. An important criticism of consistency rationality is that its accomplishment demands more cognitive abilities than any agent possesses. Especially the combination of the axioms of completeness and transitivity is extremely demanding, as it requires that agents can rank all possible alternatives consistently. Agents can rank alternatives consistently in limited and comprehensible contexts. For example, when choosing an insurance policy from a predetermined set of four policies, agents can reach a complete and transitive preference ordering. However, when they are required to rank all possible insurances in the world, including their respective attributes and possible combinations, the combination of completeness and transitivity is completely unattainable.

Assumption of complete preferences. Consistency rationality assumes that agents have complete preferences (axiom 2). This assumption is false due to three reasons. First, because of limited cognitive demands, it must be acknowledged that the completeness axiom (axiom 2), is unattainable in many situations. We defined this axiom as: for any two given alternatives x and y, an agent must either prefer the x over y, y over x, or be indifferent between the two (Hausman & McPherson, 2009, p. 5). This constraint on preferences is crucial for revealed preference theory and expected utility theory because without complete preferences, agents may be indecisive, could be inconsistent, and the theories of revealed preference and expected utility are severely limited in scope and power. Failing completeness, consistency rationality risks indeterminacy as it cannot prescribe any determinate course of action as optimal (Elster, 1991).

Preferences as unmoved movers. From the recognition of incomplete preferences, it follows that preferences must sometimes be formed. However, neoclassical economics has very little to say about how preferences are formed.

Preferences are often taken to be “stable” (Hausman, 2011, p. 16), or “exogenously given” (Dold & Schubert, 2018, p. 231) which effectively means neoclassical economists take given preference-sets as given. As Elster formulates it, preferences are “unmoved movers” (Elster, 1991, p. 21).

As a result, neoclassical welfare economics has little to contribute regarding questions how character and preferences are formed, how tastes and values change, or how economic institutions and policies affect preferences. More generally, it has been argued that neoclassical welfare economics is unable to provide a reasonable explanation of preferences that either change or internally conflict. As Sen argues it is “difficult to distinguish between inconsistencies and changing tastes” in neoclassical economics (Sen, 1977, p. 325). Likewise, neoclassical welfare economics cannot account for situations in which agents experience tension between different preferences. As a result, welfare as preference satisfaction is a problematic conception of welfare.

4.3 Behavioural welfare economics

Behavioural welfare economists have identified systematic deviations from consistency rationality. These behavioural inconsistencies cast doubt on the positive component of consistency rationality and thereby on the neoclassical welfare conception. Within behavioural welfare economics, a consensus has emerged on how to conceptualize welfare instead (Sugden, 2018). This welfare conceptualization is the dominant one in the literature on the ethics of nudging. It is based on the satisfaction of what has been called purified preferences. In this section, we argue it behavioural welfare based on preference purification is problematic.

4.3.1 Rationality and behavioural inconsistencies

Behavioural economists, most prominently Amos Tversky, Daniel Kahneman, and Richard Thaler, the last two of whom received the Nobel prize in economics, are dedicated to demonstrating that agents frequently exhibit

behaviour that deviates from consistency rationality (Tversky & Kahneman, 1974; 1981; Thaler, 2015; 2016). In this field, typically experiments are conducted in which agents are presented with a choice. It is then documented how agents systematically deviate from the stipulations of consistency rationality, by violating one of its four constraints: 1) utility maximization, 2) completeness, 3) transitivity, and 4) context-independence) (Wheeler, 2020).

1) Utility maximization. Consider the following examples of such deviations. There are experiments in which people fail to maximise expected profit in games, or financial decisions, violating utility maximization (axiom 1) (Lichtenstein & Slovic, 1971). One important factor that systematically prevents agents from maximising expected utility is the aversion of risk (also referred to as loss aversion) (Tversky & Kahneman, 1991). For example, a gamble with a 50% chance to lose €100 and a 50% chance to win €200 has a positive expected value of €50. Agents who systematically refuse such gambles because they are risk averse forego opportunities to maximise expected utility.

2) Completeness. Second, Kahneman et al. (1982) describe situations in which agents assign wrong probabilities to the occurrence of events, casting doubt over the completeness axiom (axiom 2) as systematic inability to assess probabilities leads to wrong preferences over gambles. Kahneman et al. argue that agents not only make mistakes but do so systematically, pointing to pervasive effects such as insensitivity to sample size, misconceptions of chance, insensitivity of the phenomenon regression to the mean (extreme values in a random variable are likely to be followed by more moderate values, closer to the average), and the conjunction fallacy (when agents mistakenly believe that specific conditions are more likely than broader ones.) (Kahneman et al., 1982)

3) Transitivity. Yet other experiments discovered instances in which participants reveal time-inconsistent preferences, violating transitivity (Thaler, 1981). For example, when agents are offered a choice between receiving €50 today and €60, one month from now, many people choose €50 now. When they

are offered a choice between receiving €50 in 15 months, or €60 in 16 months from now, they choose the bigger reward. This violates transitivity because in the first scenario agents prefer €10 to waiting one month, whereas in the second scenario they prefer waiting one month to €10.

4) Context-independence. The decoy effect also leads to a violation of the context-dependence (axiom 4) which states that whether an agent prefers x over y, is not affected by other alternatives becoming (un)available and that an agent's preferences should not depend on the description of the available alternatives. Yet, the description of alternatives has been shown to impact choices (Tversky & Kahneman, 1981). Other prominent examples which illustrate that aspects of the context that are influential include defaults and framing effects, creating inconsistency in choice behaviour (Tversky & Kahneman, 1981). Framing and the violation of context-independence is related to violations of the other axioms. For example, to utility-maximisation since framing alternatives in terms of losses or gains has been shown to impact decision making.

Interpretation of the inconsistencies: errors. The behavioural inconsistencies undermine the positive dimension of consistency rationality as they prove that the assumption that agents adhere to consistency rationality fails in many situations. In the words of Thaler and Sunstein, it is a "false assumption" that most people make choices that are in their best interest (Thaler and Sunstein, 2003a, p. 175). Thaler and Sunstein interpret the behavioural inconsistencies as evidence that agents are irrational, as they violate the axioms of consistency rationality. As a result, agents frequently make errors or "inferior choices" (Thaler & Sunstein, 2003, p. 175). Errors must be understood as decisions that fail to optimise one's welfare (Sunstein & Thaler, 2003a). So, an error occurs when an agent faces a choice over a set of available alternatives and chooses the alternative that does not have the highest expected utility. In conclusion, the behavioural findings falsify the positive dimension of consistency rationality and thereby revealed preference theory and expected utility theory.

Three limitations. Behavioural economists attribute the errors to a lack of proper decision-making capabilities. Instead of forming consistent preferences over all options, agents rely on heuristics to make decisions (Kahneman & Tversky, 1974). These heuristics are systematically biased, meaning agents deviate from rational standards in consistent and predictable ways due to inherent cognitive biases.

Agents are forced to rely on heuristics because the lack the ability to process all available information and compare all possible alternatives. Thaler and Sunstein define inferior choices as choices which agents "would change if they had complete information, unlimited cognitive abilities and no lack of willpower" (Thaler and Sunstein, 2003a, p. 175). They argue that if people were to possess an unlimited amount of these three factors, they would reveal consistent preferences, make perfect decisions, and fully optimise their welfare.

By discarding revealed preferences, welfare economists are faced with the challenge to formulate a new welfare conception, one that does justice to behavioural inconsistencies.

4.3.2 Purified preferences

Attempts to meet the challenge have led to the emergence of a new field, behavioural welfare economics. According to Robert Sugden, a "broad consensus" appears to be forming within behavioural welfare economics (Sugden, 2018, p. 13). This consensus consists in rejecting revealed preferences and replacing it with an alternative form of preferences, labelled "laundered preferences" by Hausman & McPherson (2009, p. 11), "true preferences" (Bershears et al., 2008, p. 3) or "purified preferences" (Infante et al., 2016, p. 1). We adopt the latter terminology. The satisfaction of these purified preference is taken to be the new welfare conception (Sugden, 2018).

Thaler and Sunstein (Sunstein & Thaler, 2003a; 2003b; 2008; Thaler 2015; Sunstein, 2014; 2015; 2016; 2017) are the most prominent voices in this

debate. Their arguments have formed the basis on which other behavioural welfare economists have created technical methods for preference purification (see Bershears et al., 2008; Bernheim 2009; 2021, Bernheim & Rangel, 2007; 2009; Bernheim et al., 2021; Salant & Rubinstein, 2008).

Purified preferences are those preferences that agents would have, if they were completely rational, rationality here understood as possessing unlimited information, cognitive abilities, and willpower. This means that proponents of preference purification must "[reconstruct] the preferences that the individual would have revealed in the absence of error" (Sugden, 2018, p. 13). These authors take actions to no longer reflect preferences but a "combined influence of true preferences and decision-making errors" or biases (Bershears et al., 2008, p. 3). This means that to discover the true preferences of agents, behavioural welfare economists can interpret agents' actions as being partially reflective of their true preferences. They only need to filter out the biases which result in decision-making errors to purify agents' actions. If this purification process is successful, they are left with the preferences agents would have revealed if they had been rational. Various methods for purification have been proposed. Sugden provides an excellent overview and summary of these methods (see Sugden, 2018, pp 52-56).

Despite rejecting the positive claims of neoclassical economics, behavioural welfare economics maintains two key normative assumptions from neoclassical theory. First, both approaches assume preferences must be complete and consistent to serve as a basis for welfare. Second, both focus exclusively on preference satisfaction rather than preference formation. While neoclassical economics assumes preferences are already complete and consistent, behavioural economics aims to reconstruct them through purification. Neither approach provides an account of how preferences are shaped or developed.

4.3.3 Welfare conception

In behavioural welfare economics, welfare is conceptualized as the satisfaction of purified preferences. This welfare conception shares with neoclassical welfare a commitment to the normative component of consistency rationality. In Thaler's words:

> Remarkably, von Neumann and Morgenstern proved that if you want to satisfy these axioms (and you do), then you must make decisions according to their theory. The argument is completely convincing. If I had an important decision to make—whether to refinance my mortgage or invest in a new business—I would aim to make the decision in accordance with expected utility theory, just as I would use the Pythagorean theorem to estimate the altitude of [a] triangle. Expected utility is the right way to make decisions. (Thaler, 2015, p. 29)

So, behavioural economics discarded the positive component of consistency rationality based on the behavioural inconsistencies but maintained the normative component. Inconsistent choices are interpreted as mistakes which harm welfare. Purified preferences are deemed to be consistent and complete. Satisfaction of purified preferences is equated with welfare. This welfare conception has become dominant in the literature on the ethics of nudging.

4.3.4 Policy implications

Behavioural welfare economics takes agents to be error-prone, which leads to the implication that policy makers should provide more assistance to agents for them to reach welfare. As a result, behavioural welfare economics has starkly different views on consumer sovereignty and paternalism compared to neoclassical welfare economics. Whereas neoclassical welfare economics places emphasis on consumer sovereignty in many domains and is strongly suspicious

of paternalism, with behavioural economics this is the other way around (Jolls et al., 1998, p. 1543). Behavioural welfare economists believe consumers need more guidance and protection when making economic decisions.

It has been argued that the government can provide such guidance by nudging people towards welfare.[8] The argument in favour of benevolent governmental nudging almost flows naturally from the notion of purified preferences. Since 1) agents frequently make errors, limiting their welfare; and 2) revealed preferences are context dependent; it follows that changing the contexts of agents so they make better decisions would lead to more welfare. Simply stated: the government should nudge people towards welfare. As shown in chapter 2, this argument has received ample attention in the debate in the ethics of nudging. Moreover, benevolent governmental nudging is not merely a theoretical idea as it has relevant applications in the real world and is quickly becoming accepted in policy circles (see for example Halpern, 2015). Despite its focus on consumer protection, behavioural welfare economists have so far not extensively discussed the potential welfare harms from private nudges.

4.3.5 Critical evaluation

Conceptualizing welfare as the satisfaction of purified preferences has several advantages. First, it "grants that people may prefer something that is bad for them because of ignorance or cognitive defects" (Hausman & McPherson, 2009, p. 11). Many people have the experience of doing something and later regretting it, or of acting based on faulty premises and later learning what the correct premises were. A theory which can account for such experiences is a step in the right direction. In other words, behavioural welfare accounts for limits in information, cognitive abilities, and willpower.

[8] The idea that governments should use nudges to optimise welfare in the behavioural economic sense is often referred to as "libertarian paternalism" (Sunstein & Thaler, 2003b, p. 175).

Second, behavioural welfare allows for choices to be impacted by the context or the environment. Empirical evidence shows that agents change their choices when choosing under different circumstances and that actors are influenced by the context in which they make decisions. In other words, behavioural welfare can account for the concept of nudging, whereas neoclassical welfare cannot. However, there are important problems with this welfare conception.

Inner rational agent. Behavioural welfare assumes agents have consistent and complete preferences, which do not exist. Thereby, it postulates an "inner rational agent" (Infante et al., 2016). This means that behavioural welfare supposes agents have a "latent capacity, constant across decision environments, to form context-independent subjective judgements on the basis of error-free reasoning" (Sugden, 2015, p. 6).

Thaler and Sunstein state government nudges can "make choosers better off, as judged by themselves". The 'themselves' refers the nudged agent. According to Thaler and Sunstein, this agent can reflectively endorse the nudge (Thaler & Sunstein, 2008, p. 5). Such reflective endorsement hinges on the existence of consistent and complete preferences.

Furthermore, without complete and consistent preferences, the notion of purification becomes meaningless as this would mean there is nothing to extract through purification. Preference purification theorists can deny postulating an inner rational agent, but this would only cast more doubt over their welfare conception. The process of purification needs some form of raw material to purify. Otherwise, the behavioural welfare conception cannot be a form of preference satisfaction, as its advocates maintain.

Preferences as unmoved movers. As behavioural welfare incorporates an inner rational agent, it incorporates the same problem that neoclassical welfare

faced, namely it fails to provide an account of preference formation. Purified preferences are assumed to exist, yet there is no idea of how these are formed. This problem is the same as the neoclassical version of the problem, which we discussed earlier.

These problems stem directly from behavioural welfare's inheritance of two problematic assumptions from neoclassical economics: the requirement for complete and consistent preferences, and the exclusive focus on preference satisfaction rather than formation. By maintaining these assumptions while rejecting consistency rationality as a descriptive theory, behavioural welfare creates an internal contradiction: it acknowledges that actual preferences are neither complete nor consistent yet requires reconstructing preferences that have exactly these properties. This contradiction manifests in the problematic notion of an 'inner rational agent' (Infante et al., 2016).

Excessive paternalism. Behavioural welfare is contingent on preference purification, which easily surmounts into excessive paternalism. Once it is accepted that people frequently make errors and are faulty judges of what brings them welfare, it is a small step to imposing actions and policies on them for their own good. Many behavioural welfare economists argue the government must nudge agents because they make poor welfare choices. This strategy has however been critiqued for being excessively manipulative, violating autonomy and showing a lack of respect (see chapter 2). Governmental nudging neglects that people derive value and learnings from their mistakes. It focuses solely on realising outcomes yet dismisses the value of the process.

Preferences purification and nudges surmount into excessive paternalism (Whitman & Rizzo, 2007). Preference purification hinges on the assessment that it is easier for policy makers and preference purification theorists to determine an agents' true preferences, than it is for the agents themselves. This assumption is questioned by Rizzo & Whitman, who illustrate with a quote of Bentham: "It is

a standing topic of complaint, that a man knows too little of himself. Be it so: but is it so certain that the legislator must know more?" (Rizzo & Whitman, 2009, p. 905). Purification of preferences is an "expectedly contentious" process which easily leads to the classification of socially desired behaviour as true preferences and of socially undesired preferences as mistaken preferences. (Hansson & Grüne-Yanoff, 2018, p. 37).

4.4 Behavioural welfare without purified preferences

4.4.1 Three fundamental limitations

Behavioural welfare economics acknowledges the three fundamental limitations of information, cognitive abilities and willpower, yet refuse to take these limitations into their analysis. Instead, it advocates working around the limitations by purifying preferences under a hypothetical ideal scenario in which the limitations are absent, which causes the problems described above. In our view, behavioural welfare economics should not focus on purifying preferences to reconstruct what agents would prefer under ideal circumstances. Instead, it should accept and embrace the limitations as the starting point of its analysis.

If we take seriously the three limitations, we must acknowledge complete and consistent preferences cannot be found in agents. As March notes, limitations in cognitive ability "affect information processing about preferences just as they affect information processing about consequences. [...] Human beings have unstable, inconsistent, incompletely evoked, and imprecise goals at least in part because human abilities limit preference orderliness" (March, 1978, p. 598). Dynamic interaction and uncertainty further contribute to this incompleteness.

Since limitations in information, cognitive ability, and willpower render consistent and complete preferences unattainable, welfare based on preference satisfaction faces a fundamental problem since the satisfaction of incomplete or

inconsistent preferences is no guarantee for welfare. This problem was recognized by Nobel laureate Herbert Alexander Simon (1957), who coined the term bounded rationality to describe what we have called the three fundamental limitations. Bounded rationality is concerned with describing how agents can function optimally, within their limitations. They do so by seeking satisfactory options, rather than searching for utility maximizing options.

In Simon's perspective, decisions made within the boundaries formed by the fundamental limitations can be rational and appropriate, considering the limitations agents face. This step means letting go of outcome rationality (which is referred to as substantive rationality by Simon). Outcome rationality evaluates decisions based on whether they achieve optimal results (Simon, 2001). Instead, this approach focuseson the processes an agent can improve in order to improve their preferences and decision-making capabilities (Simon, 2001). We agree with Simon that “there is a close affinity between bounded and procedural rationality” (p. 12783) Acknowledging rationality is bounded shifts the focus from the evaluation of optimal outcomes to the description of optimal processes.

Recognizing that preferences are necessarily incomplete and inconsistent, also highlights the need for a theory of preference formation and adjustment (c.f. Mckenzie et al., 2018; Dold & Schubert, 2018; Fabian & Dold, 2022). Where preferences are incomplete, they must be shaped. Where they are inconsistent, they must be adjusted. We use the term preference shaping to encapsulate this twofold process of shaping and adjusting existing. The quality of these processes, rather than the consistency of preferences at any moment, becomes a crucial dimension of welfare. Importantly, welfare still involves the satisfaction of preferences but must be complemented with their proficient shaping.

This approach connects welfare to the practical realities of human decision-making. In addition, it offers a framework for evaluating digital nudges based on how they affect our ability to both shape and satisfy preferences through their effect on information, cognitive abilities, and willpower. In the

coming chapters, we argue digital nudges hurt welfare through their effect on the three fundamental limitations. In the remainder of this chapter, we analyse the mechanisms required for preference shaping.

4.4.2 Preference shaping

The task at hand is to formulate what Valerie Tiberius calls “a first-personal, process-based account of how to live, as opposed to an impersonal, goal-based theory of the good life” (Tiberius, 2008, p. 3). The central question is not which life is objectively best, but how agents ought to shape their own preferences over time. This question becomes urgent once we acknowledge two features, emphasized in this dissertation: first, agents face three fundamental limitations; second, digital environments subject them to pervasive and potentially manipulative nudges that influence not only their choices but the formation of their preferences themselves.

An adequate normative account of preference shaping must explain when such shaping is autonomous and when it is not. The issue is therefore not merely welfare, but authority over the development of one’s preference structure. We require an account that integrates welfare with a rich conception of autonomy. In terms of chapter 2: not autonomy conceptualized as freedom of choice, but autonomy conceptualized as self-constitution (Vugts, 2020). This view conceptualizes agents as authors of their own preference structures over time.

For this reason, we turn to John Stuart Mill’s account of individuality and personal development (Mill, 2001/1859). Mill’s view is particularly instructive because it combines a subjective and pluralistic conception of welfare with a developmental understanding of autonomy. For Mill, there is no universal blueprint of the good life. Individuals must discover for themselves which ways of living contribute to their welfare. This discovery is not passive; it requires active self-direction. As Mill famously argues, "the free *development”* of one’s preferences as “one of the leading essentials of well-being" (Mill, 2001/1859, p.

53; emphasis mine). This development occurs through a deliberate process of self-discovery and preference shaping through experience and reflection.

Mill thus links welfare to a process of deliberate preference formation. An agent whose "desires and impulses are not his own, has no character, no more than a steam-engine has a character" (2001/1859, p. 56). Interpreted in contemporary terms, character can be understood as one's preferences set. We agree with McPherson, who argues that "Mill's theory can perhaps best be seen as an attempt to analyse the question of how one would want one's character or system of preferences to develop over time-or, more abstractly, how one would go about choosing between two different preference systems for oneself" (McPherson, 1982, p. 260).

Autonomy, on this view, requires that this set emerges through self-directed experimentation rather than uncritical absorption of external influences. The core of Mill's account is therefore developmental. Agents must engage in a process of forming and revising their preferences through experience and reflection.

Building on this Millian foundation, Fuchs (2001) offers a contemporary view on Mill and suggests elements that we can take as the core for proficient preference shaping. We suggest that proficient preferences shaping involves (1) engaging in experiments in living in order to discover which activities and values genuinely contribute to one's welfare, and (2) cultivating evaluative and reflective capacities that enable agents to assess, compare, and revise their preferences over time. These components capture the practical and deliberative dimensions of preference shaping that are essential if preferences are to be shaped autonomously and proficiently.

1) Experiments in living. Preference shaping happens through "experiments in living" (Mill, 2001/1859, p. 53). Mill uses this term to express that agents should test different ways of life. Experimenting is essential, as it

allows agents to discover which lifestyles yield the greatest individual welfare. As welfare is individually subjective, agents must experiment, so they acquire various experiences and can identify which of those are appropriate to their way of living. Tradition and custom should be treated with a sense of suspicion. According to Mill, "there is only too great a tendency in the best beliefs and practices to degenerate into the mechanical; and [only] a succession of persons [with] ever recurring originality prevents the grounds of those beliefs and practices from becoming merely traditional" (2001/1859, p. 60). Proficient preference shaping requires discovering rather than mindlessly adopting.

For experimentation to successfully contribute to proficient preference shaping, agents must not only collect experiences, but they must also learn to evaluate their experiences. By experimenting, agents develop "[t]he human faculties of perception, judgment, discriminative feeling, mental activity, and even moral preference" (Mill, 2001/1859, p. 54). The idea is that by repeatedly choosing activities and styles that have potential merit and by skilfully assessing their merit, agents come to understand what gives them welfare. Experimentation is thus a skill that must be learned.

*2)Reflective capacities***.** Proficient preference shaping is impossible without evaluative and reflective capacities to compare and understand different preferences. To better understand this reflective capacity, we turn to Frankfurt's concept of meta-preferences. The aspiration to better preferences implies a decision amongst various preference sets. Such a decision between preference sets can only be made based on a meta-preference. This meta-preference indicates a reflective endorsement of one's preferences, pointing to a deeper level of self-awareness and deliberative capacity in personal development.

The notion of meta-preferences originates from Harry Frankfurt's hierarchical account of the will, which offers a framework for analysing the reflective capacities required for proficient preference shaping. Frankfurt

distinguishes between “first-order” (1971, p. 7) and “second-order desires” (p. 6). We use the term preferences instead of desires First-order preferences are basic preferences over choice alternatives. Second-order preferences are preferences over preferences. Humans may prefer to prefer different things than they actually prefer. Second-order preferences must be distinguished from “second-order volitions” (p. 10). These reflective endorsements of certain preferences that move an agent to act. Frankfurt’s central insight is that agents do not merely have preferences; they can take a stand on them. This hierarchical structure introduces a normative dimension internal to the agent’s motivational system. Some preferences are not simply experienced; they are reflectively endorsed as authoritative.

This hierarchical model clarifies why reflective capacities are indispensable for proficient and autonomous preference shaping. To shape one’s preferences over time is to decide, at a higher level, which desires should guide one’s conduct and which should be resisted or revised. Such decisions presuppose meta-preferences, which can be thought of as preferences concerning one’s preference set. Without this capacity for higher-order evaluation, the development of one’s preferences would be merely the outcome of causal forces, whether social, environmental, or manipulative. In Frankfurt own words, without second-order volitions, agents cannot shape their preferences due to “lack of capacity for reflection or [...] mindless indifference to the enterprise of evaluating [one’s] own desires and motives” (1971, p. 13). Frankfurt’s framework thus supplies the conceptual bridge between autonomy as self-constitution and the governance of one’s evolving preference structure. Reflection enables agents to compare possible configurations of their motivational system and to endorse one as expressive of who they take themselves to be. Therefore, reflection is a necessary condition for authority over the shaping of one’s preferences and, consequently, over one’s welfare.

Having conceptualized proficient preference shaping as a process of experimenting and reflecting, we must highlight how this process is attainable for cognitively limited agents, as opposed to agents possessing full or idealized rationality. This is consistent with Tiberius, who proposes "a different strategy", namely "to think about how to train the rational and reflective capacities we actually have so that they can function together with our emotions, moods, and desires to get us somewhere we'd like to be" (2008, p. 7). This strategy is purposefully modest; it is attainable for humans facing the three fundamental limitations. Reflection for Tiberius is not an idealized stance. Reflection is a basic human capacity that they engage in throughout their lives to give their life direction in a way that is meaningful to them. At the same time, this approach recognizes the value of non-reflective emotions and desires. Formulated in terms of Haidt, as we did in the introduction, this acknowledges both the elephant and the rider. "To do this, we need to think about standards for improvement that take seriously the ways in which explicit reflection tends to go wrong and the ways in which non-reflective experience can lead us in the right direction, [in ways] that are useful for us, given our needs and limitations." (Tiberius, 2008, p. 7).

Before continuing, we should emphasize that Tiberius succeeds in opening the room for a normative account that acknowledge the fundamental limitations by "beginning with the first-person question about how to live [which] results in an account that addresses people on their own terms and, therefore, has a foothold in people's practical reasoning" (Tiberius, 2008, p. 14). In a behavioural world, where the core insight is that people are not as rational as was long expected, the best guidance is not to find more clever ways of defining the target (as outcome rationality suggest). Rather, "the best guidance is guidance we can follow" (2008, p 14).

4.4.3 From process rationality to welfare: how the three fundamental limitations impact preferences

While outcome rationality is distorted by inconsistency and incompleteness, in process rationality, inconsistency and incompleteness serve as critical factors that actively mould an individual's values and tastes. As Rizzo & Whitman (2018) argue, the behavioural inconsistencies which are classified as errors by preference purification theorists may be seen as a positive force for learning when taking a more dynamic perspective to decisions over time. Making errors allows agents to try different approaches, gain experience, and develop their values and tastes. Confronting one's incomplete and inconsistent preference set with the world through making choices and acting in it is like holding a blade against a stone to sharpen it. By making errors in the world, agents come to understand their inconsistencies and incompleteness preferences over time. Through continuously engaging in this cycle of action and reflection, an iterative sharpening process arises so that, over time, experienced agents gravitate towards better preferences as their comprehension of their meta preferences deepens. For example, through sampling of different sports, agents may not only discover or develop a taste for one, but they may also discover what their core values are, be it health, leisure, or friendship. In this way, experimentation not only leads to more stable and fulfilling preferences but also to a development of one's hierarchy of preferences. Through dealing with inconsistencies and incompleteness and by experimenting into the unknown, agents learn "to perceive the interconnectedness of [...] interests so as to pursue multiple ends harmoniously and coherently" (Fuchs, 2001, p. 237). The metaphor of the stone and the blade applies here as well; agents become aware of their gaps in knowledge and mistaken beliefs only by engaging with the world.

Having established how process rationality contributes to preference shaping, we can now describe welfare more precisely in relation to this dynamic

conception. Welfare is then the satisfaction of preferences that are shaped through experimentation and reflection. Crucially, the shaping must occur under conditions that preserve the agent's capacity for experimentation and higher-order reflection. Both are influenced by three fundamental limitations that all agents face: imperfect information, bounded cognitive abilities, and limited willpower. Understanding how these limitations affect both preference shaping and preference satisfaction provides important insights for our conception of welfare.

Information. Information aids proficient preferences shaping because it allows agents to understand their options. While information is difficult to define precisely, as we discuss in chapter 5, functionally we can understand information as the input for decisions. To shape their preferences, agents need information about the available alternatives. In addition, making choices provides them with informational feedback which helps them to judge which preferences are worth pursuing. When they conduct experiments, they learn their preferences through their emotional reactions. Fabian & Dold (2022) call this "affective feedback" (p. 9).

Agents also need information to understand how their actions contribute to the satisfaction of their preferences. For example, if agents prefer to improve their diet by limiting their sugar intake, they need information on which products contain sugar and where they can find these products. It also helps to have information on which synonyms of sugar are typically used on food packaging (such as dextrose).

When information is particularly distorted, coloured, or incomplete, both preference shaping and satisfaction become compromised. Agents may develop preferences based on misconceptions or fail to effectively pursue their genuine preferences.

Cognitive abilities. Cognitive abilities aid preference shaping because they allow agents to reflect on their available information, experiences and preferences. While information is the input for decisions, cognitive abilities can be thought of as the processing of information. Reflection is crucial to detect inconsistencies between various preferences, understand how preferences relate to meta-preferences, and shape preferences over time, as we discuss in chapter 6. As Mill emphasizes, reflective and evaluative capacities enable agents to compare different experiences and develop their preferences. These cognitive abilities help agents process their experiments in living, allowing them to learn meaningful insights about themselves and what they prefer.

Cognitive abilities also help to satisfy preferences because understanding cause-effect relationships and reasoning allows agents to determine how they must act to satisfy their preferences. Once agents have shaped their preferences through reflection and experimentation, they need cognitive capacities to translate these preferences into practical action. This involves identifying appropriate means to achieve their ends, anticipating obstacles, and adjusting strategies accordingly. For example, if agents prefer to improve their diet, they require cognitive abilities to plan meals, understand various macronutrients, and understand how this preference relates to other preferences such as preserving spontaneity. In addition, it requires cognitive abilities to judge the merit of the many health claims and advice from (semi)experts that many dieters encounter.

When cognitive abilities are especially limited, agents struggle to integrate their experiences into a coherent preference set, often failing to recognize patterns or inconsistencies in their preferences. Agents with severely limited cognitive abilities may be unable to shape autonomous preferences at all, relying instead on social imitation or external guidance alone.

Willpower. Preference shaping requires a continuous dedication to experimentation and reflection, and willpower aids in maintaining this

commitment. As we discuss in chapter 7, even when agents have the necessary information and cognitive abilities to make sound decisions, they may fail to implement and act on their shaped preferences. Hence, willpower is necessary for the execution of decisions. It aids the output of the decision-making process. Due to familiarity bias, agents face temptations to postpone experimenting and stick with the familiar. Due to present bias, agents face temptations to indefinitely postpone experiments in living. Developing one's higher-order preferences often requires sustained effort and initial discomfort. For example, an individual may wish to cultivate a preference for intellectual challenge rather than passive entertainment, even though they currently feel a stronger pull toward familiar and effortless forms of leisure. Developing this higher-order preference requires repeatedly choosing demanding activities despite initial resistance and discomfort. Willpower is needed to sustain such efforts. Finally, confirmation bias tempts agents to skip reflection in favour of selectively attending to information that validates their existing preferences.

Willpower enhances an agent's ability to satisfy their preferences by enabling them to remain committed to their higher-order preferences, which they have reflectively endorsed. For example, an agent may endorse a higher-order preference to become someone who values moderation in diet, even though they continue to experience strong first-order desires for sugary foods. Each visit to a supermarket presents a temptation to act contrary to this higher-order commitment. Willpower, expressed in anticipating such temptations and deliberately structuring one's environment or routines to manage them, is essential for aligning action with one's reflectively endorsed higher-order preferences.

When willpower is particularly limited, agents frequently act contrary to their higher-order preferences and instead satisfy immediate impulses, despite their reflective judgment that these actions do not align with their welfare. This

creates a disconnect between what agents believe would contribute most to their welfare and their actual behaviour.

This framework of process rationality and the three fundamental limitations has significant implications for policy. Rather than identifying citizen preferences, policy makers should instead focus on creating conditions that enable agents to shape their preferences proficiently.

4.4.4 Policy implications

A focus on shaping preferences releases policymakers and governments from the need to identify preferences. This alleviates concerns regarding heterogeneity, epistemic access and privacy (discussed in chapter 2) as well as worries around incomplete preferences and excessive paternalism hurting autonomy (discussed earlier in this chapter). Instead of identifying citizen preferences, policy makers should focus on creating the conditions in which agents can function well, given the three limitations. This means, governments must focus on creating conditions in which agents can shape and satisfy their preferences. Most importantly, this requires creating the conditions in which agents can improve their information, cognitive abilities, and willpower. The free market, freedom of speech, functioning democracy, and free press are all mechanisms that ensure information flows freely. Protecting these mechanisms is the most important thing governments can do to ensure agents have access to quality information. To improve cognitive abilities, governments must facilitate education that helps students reflect and reason proficiently. Finally, to aid willpower, governments must allow citizens to make their own decisions, as willpower requires practicing making important decisions for oneself repeatedly.

This leaves governments and policy makers with three options with respect to behavioural interventions. First, they can refrain from intervening in the behaviour of citizens. Second, they can justify nudges on other normative

standards. For example, one could argue it is warranted for the government to nudge citizens into climate friendly behaviour, despite their preferences. Third, governments can use a different form of behavioural intervention, based on processes rather than outcomes, aimed at improving the information or cognitive abilities (like reasoning) of citizens. Unlike nudges, which are justified by depicting agents as error-prone decisionmakers and by pushing agents towards better outcomes, these interventions are justified by improving reasoning, so that agents are placed in a position where they can choose well for themselves. There are three types of interventions. First, boosts, which seek "to improve people's decision-making competences by enriching their repertoire of skills and decision tools or by restructuring the environment such that existing skills and tools can be more effectively applied" (Barton & Grüne-Yanoff, 2015, p. 349) (see Grüne-Yanoff and Hertwig, 2016; Hertwig and Grüne-Yanoff, 2017; Grüne-Yanoff, 2018). Second, debiasing interventions, which stimulate active thinking, aiming to engage robust cognitive processes rather than the shallow cognitive processes targeted by nudges (see Croskerry et al., 2013; Isler et al., 2020). Third, nudge+ interventions, which are hybrid behavioural interventions that encourage agents to think about their choice in combination with a nudge (Banerjee *et al.*, 2023) (see Banerjee and John 2021; 2022; 2023).

Policy makers should focus on creating beneficial conditions that aid information, cognitive abilities, and willpower. Doing so will improve welfare policies in general and mitigate the adverse welfare effects of digital nudges in particular.

4.5 Conclusion

In this chapter, we have critiqued both neoclassical and behavioural welfare economics, particularly examining how the latter maintains elements from its predecessor despite recognizing cognitive limitations. While

behavioural welfare economics acknowledges the three fundamental limitations—imperfect information, bounded cognitive abilities, and limited willpower—it still clings to consistency rationality through preference purification.

Our framework proposes a shift from consistency or outcome rationality to process rationality. Rather than attempting to salvage consistency through preference purification, we embrace the inherent inconsistency and incompleteness of preferences as integral to their development over time. This framework recognizes inconsistencies not as deviations to be corrected, but as opportunities for learning and preference shaping through real-world engagement.

Process rationality centres on how agents can optimally navigate decision-making given their fundamental limitations. We argue that welfare emerges from two interdependent processes: preference shaping and preference satisfaction. Drawing from Mill, we identify three critical components of proficient preference shaping: experiments in living, and cultivation of reflective capabilities.

The implications of this framework extend beyond theoretical understanding to practical policy applications. Rather than identifying and satisfying static preferences, policymakers should focus on creating environments that facilitate agents' ability to shape and satisfy their preferences effectively considering their fundamental limitations. The following three chapters will examine how digital nudges specifically affect each fundamental limitation, thereby influencing welfare through their impact on preference shaping and satisfaction.

4.6 Bibliography

Arrow, K. (1970). *Essays in the Theory of Risk Bearing*. Amsterdam: North Holland.

Atkinson, A. B. (2001). The strange disappearance of welfare economics. *Kyklos*, 54(2-3), 193-206.

Banerjee, S., & John, P. (2021). Nudge plus: Incorporating reflection into behavioral public policy. *Behavioural Public Policy*, 1–16. https://doi.org/10.1017/bpp.2021.6

Banerjee, S., & John, P. (2024). Nudge plus: incorporating reflection into behavioral public policy. *Behavioural Public Policy*, 8(1), 69-84.

Banerjee, S., Grüne-Yanoff, T., John, P., & Moseley, A. (2023). It's Time We Put Agency into Behavioural Public Policy. *Available at SSRN 4325117*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4325117

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

Bernheim, B. D. (2009). Behavioral welfare economics. *Journal of the European Economic Association*, *7*(2–3), 267–319.

Bernheim, B. D. (2021). In defense of behavioral welfare economics. *Journal of Economic Methodology*, *28*(4), 385–400.

Bernheim, B. D., Braghieri, L., Martínez-Marquina, A., & Zuckerman, D. (2021). A theory of chosen preferences. *American Economic Review*, *111*(2), 720–754.

Bernheim, B. D., & Rangel, A. (2007). Toward choice-theoretic foundations for behavioral welfare economics. *American Economic Review*, *97*(2), 464–470.

Bernheim, B. D., & Rangel, A. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *The Quarterly Journal of Economics*, *124*(1), 51–104.

Beshears, J., Choi, J., Laibson, D., & Madrian, B. (2008). How are preferences revealed? *Journal of Public Economics, 92*, 1787–1794.

Brandt, R. B. (1979). *A theory of the good and the right*. Oxford University Press.

Burns, T. (2016). "Happy slaves"? The adaptation problem and identity politics in the writings of Amartya Sen. *International Journal of Social Economics*, *43*(12), 1178–1193.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality & Safety*, 22(Suppl 2), ii58-ii64.

Dold, M. (2023). Behavioural welfare economics: Foundations, approaches and trends. *Fiscal Studies*, *44*(2), 137–150. https://doi.org/10.1111/1475-5890.12340

Dold, M. F., & Schubert, C. (2018). Toward A Behavioral Foundation of Welfare economics. *Review of Behavioral Economics*, *5*(3–4), 221–241. https://doi.org/10.1561/105.00000097

Elster, J. (1991). When rationality fails. In K. S. Cook & M. Levi (Eds.), *The limits of rationality* (Vol. 19). University of Chicago Press Chicago.

Elster, J. (2009). *Reason and rationality*. Princeton University Press.

Elster, J. (2016). *Sour Grapes: Studies in the subversion of rationality*. Cambridge University Press. https://doi.org/10.1017/CBO9781316494172

Fabian, M., & Dold, M. (2022). Agentic preferences: A foundation for nudging when preferences are endogenous. *Behavioural Public Policy*, 1–21. https://doi.org/10.1017/bpp.2022.17

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy, 68*(1), 5–20.

Friedman, M. (1953). The methodology of positive economics. In *Essays in positive economics* (pp. 3-43). University of Chicago Press.

Fuchs, A. E. (2001). Autonomy, slavery, and Mill's critique of paternalism. *Ethical Theory and Moral Practice*, *4*, 231–251.

Griffin, J. (1986). *Well-being: Its meaning, measurement, and moral importance*. Oxford University Press.

Grüne-Yanoff, T., & Hertwig, R. (2016). Nudge Versus Boost: How Coherent are Policy and Theory? *Minds and Machines*, *26*(1), 149–183. https://doi.org/10.1007/s11023-015-9367-9

Grüne-Yanoff, T. (2018). Boosts vs. Nudges from a Welfarist Perspective. *Revue D Economie Politique*, *128*(2), 209–224.

Halpern D (2015) *Inside the nudge unit: How small changes can make a big difference*. Random House.

Hansson, S. O., & Grüne-Yanoff, T. (2018). Preferences. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/preferences/

Hargreaves Heap, S. P. (2013). What is the meaning of behavioural economics? *Cambridge Journal of Economics*, *37*(5), 985–1000. https://doi.org/10.1093/cje/bes090

Harsanyi, J. C. (1977). Morality and the Theory of Rational Behavior. *Social Research*, *44*(4), 623–656.

Hausman, D. M. (2011). *Preference, Value, Choice, and Welfare*. Cambridge University Press. https://doi.org/10.1017/CBO9781139058537

Hausman, D. M. (2018). Efficacious and Ethical Public Paternalism. *Review of Behavioral Economics*, *5*(3–4), 261–280. https://doi.org/10.1561/105.00000090

Hausman, D. M., & McPherson, M. S. (2006). *Economic analysis, moral philosophy, and public policy*. Cambridge University Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=149844

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science*, *12*(6), 973–986. https://doi.org/10.1177/1745691617702496

Infante, G., Lecouteux, G., & Sugden, R. (2016). Preference purification and the inner rational agent: A critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, *23*(1), 1–25. https://doi.org/10.1080/1350178X.2015.1070527

Isler, O., Katsikopoulos, K. V., & Gigerenzer, G. (2020). How to activate intuitive and reflective thinking in behavior research. *Journal of Behavioral Decision Making*, 33(5), 569-576. https://doi.org/10.1002/bdm.2165

Jolls, C., Sunstein, C. R., & Thaler, R. H. (1998). *A Behavioral Approach to Law and Economics* (SSRN Scholarly Paper 74927). https://papers.ssrn.com/abstract=74927

Kahneman, D. (2011). *Thinking, fast and slow*. macmillan.

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

Knight, F. H. (1921). *Risk, uncertainty and profit* (Vol. 31). Houghton Mifflin.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, *89*(1), 46.

March, J. G., & Simon, H. A. (2005). Cognitive Limits on Rationality. In M. H. Bazerman (Ed.), *Negotiation, decision making and conflict management, Vol. 1–3* (pp. 201–237). Edward Elgar Publishing.

March, J. G. (1978). Bounded rationality, ambiguity, and the engineering of choice. *The Bell Journal of Economics*, 587–608.

McPherson, M. S. (1982). Mill's Moral Theory and the Problem of Preference Change. *Ethics*, *92*(2), 252–273.

McKenzie, C. R., Sher, S., Leong, L. M., & Müller-Trede, J. (2018). Constructed preferences, rationality, and choice architecture. *Review of Behavioral Economics*, *5*(3–4), 337–360.

Mill, J. S. (2001). *On liberty* (Original work published 1859). Batoche Books.

Mill, J. S. (2001). *Utilitarianism* (Original work published 1863). Batoche Books.

Mongin, P. (1998). Expected utility theory. In J. B. Davis, D. W. Hands, & U. Mäki (Eds.), *The handbook of economic methodology* (pp. 171-179). Edward Elgar Publishing.

Ng, Y.K. (2004). *Welfare Economics*. Palgrave Macmillan UK.

Railton, P. (1986). Moral realism. *The Philosophical Review, 95*(2), 163–207. https://doi.org/10.2307/2185589

Rizzo, M. J., & Whitman, G. (2018). Rationality as a process. *Review of Behavioral Economics*, *5*(3–4), 201–219.

Rizzo, M. J., & Whitman, D. (2009). The knowledge problem of new paternalism. *Brigham Young University Law Review*, 2009(4), 905-968.

Salant, Y., & Rubinstein, A. (2008). (A, f): choice with frames. *The Review of Economic Studies*, *75*(4), 1287-1296.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, 5(17), 61-71.

Schubert, C. (2015). Opportunity and preference learning. *Economics & Philosophy*, *31*(2), 275–295.

Sen, A. K. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & public affairs*, 317-344.

Simon, H. A. (1957). *Models of man; social and rational.* https://psycnet.apa.org/record/1958-00363-000

Simon, H. A. (1996). Designing organizations for an information-rich world. *International Library of Critical Writings in Economics*, *70*, 187-202.

Simon, H. A. (2001). Rationality in Society. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 12782–12786). Pergamon. https://doi.org/10.1016/B0-08-043076-7/01953-7

Smith, M. (2004). *Ethics and the a priori*. Cambridge University Press.

Sugden, R. (2018). The community of advantage: *A behavioural economist's defence of the market*. Oxford University Press.

Sunstein, C. R., & Thaler, R. H. (2003b). Libertarian paternalism is not an oxymoron. *The University of Chicago law review*, 1159-1202.

Sunstein, C. R. (2014). Why nudge: The politics of libertarian paternalism. Yale University Press.

Sunstein, C. R. (2015). Choosing not to choose: Understanding the value of choice. Oxford University Press, USA.

Sunstein, C. R. (2016). Do People Like Nudges? *Administrative Law Review*, *68*(2), 177–232.

Sunstein, C. R. (2017). Misconceptions About Nudges. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3033101

Thaler, R. H., & Sunstein, C. R. (2003a). Libertarian paternalism. *American economic review*, *93*(2), 175-179.

Thaler, R. H. (2015). Misbehaving: The making of behavioural economics. Penguin Books.

Thaler, R. H. (2016). Behavioral Economics: Past, Present, and Future. *American Economic Review*, *106*(7), 1577–1600. https://doi.org/10.1257/aer.106.7.1577

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Penguin.

Tiberius, V. (2010). The reflective life: Living wisely with our limits. OUP Oxford.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton University Press.

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

Wheeler, G. (2020). Bounded rationality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/bounded-rationality/

Whitman, D. G., & Rizzo, M. J. (2007). Paternalist slopes. *NYU Journal of Law & Liberty*, 2(3), 411-482.

Chapter 5 – How digital nudges distort information

5.1 Introduction

In this chapter, we discuss the first of the three fundamental limitations, information. We first provide a brief overview of information defined by its functional role in decision-making, establishing why this understanding is crucial for analysing how digital nudges can undermine the informational basis of welfare. Information forms the essential foundation for all decision-making and preference shaping, and in digital environments, it is particularly vulnerable to manipulation as its presentation and filtering are controlled by algorithms and design choices. We then argue digital nudges cause distorted information in two ways. First, they cause agents to be confronted with and consume an overload of information. Second, they cause agents to consume systematically biased information. We argue digital nudges systematically favour five types of information: belief-consistent, social, negative, predictive, and empty. Third, we discuss how these two information distortions hamper preference shaping. Fourth and finally, we provide recommendations on how teachers and educational policy makers can prepare students to manage their information environments to proficiently shape preferences and attain welfare.

5.2 Understanding information's function in decision-making

Information is an elusive concept. According to Floridi, it is “notorious for coming in many forms and having many meanings” (Floridi, 2010, p. 1). To avoid unnecessary complications, we define information in terms of its function. Agents use information as input for their decisions, it’s the foundation on which decisions are based. Information is the opposite of uncertainty in the sense that

that it provides a basis on which beliefs and decisions can be constructed, while uncertainty provides none (Sequoiah-Grayson & Floridi, 2022). When nothing is known about the environment, agents have no information available, and they operate in complete uncertainty.

In contrast, when the environment provides recognisable and exploitable patterns, it is possible to derive information from the environment which can then be used to behave effectively (Sequoiah-Grayson & Floridi, 2022). As such, "we commonly think of information as a means to an end" (Golman et al., 2017, p. 96). It helps us behave more effectively and increase the chances we attain our goals.

For this reason, information is generally considered to be valuable. According to Floridi:

> information has economic value because of its usefulness: it allows agents to take courses of actions (consider options, avoid errors, make choices, take rational decisions, and so forth) that normally yield higher expected payoffs (expected utility) than the agents would obtain in the absence of such information. (Floridi, 2010, p. 90)

This economic value of information explains why its control and distribution have become central to digital economies. In digital environments, the supply of information is algorithmically mediated, with substantial consequences for consumers. As argued in chapter 3, information and its consumption are personalized and measured. When users navigate digital environments, they encounter information that has been deliberately structured, filtered, and presented according to the goals of platform designers, namely capturing attention and steering users according to the economic incentives of the digital designer.

In the context of the current discussion on digital nudges, information is everything that is known concerning the available alternatives over which an agent makes a decision, which is referred to as the choice-set. Information

concerning the choice-set includes for example the qualities, costs and availability of the available alternatives. This information naturally determines how agents perceive and value the alternatives, as is well documented in neoclassical economic literature.

Nudges by definition do not impact the choice-set, nor the information concerning the choice-set (see chapter 2). However, nudges do impact the way in which a given choice-set and the information regarding that choice-set is presented, which is referred to as the choice-architecture. As the literature on behavioural economics has described, the choice architecture influences how agents perceive and react to the choice-set (see chapter 4).

Nudges can exert influence by framing pieces of information in a way that makes certain aspects more salient. For example, Tversky and Kahneman provide an example, in which two groups of correspondents are asked to make a decision regarding a fictional medical situation. The first group is presented with the following choice:

> [i]f Program A is adopted, 200 people will be saved. If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved" (Tversky & Kahneman, 1981, p. 453). Program A is favoured 72% to 28%. The second group is presented with the following choice: "[i]f Program C is adopted 400 people will die. If Program D is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. Which of the two programs would you favour? (Tversky & Kahneman, 1981, p. 453)

Program D is favoured 78% to 22%. The difference in answers between the two groups is startling, as the two problems and the respective probabilities are in essence identical. The only difference between the two problems is that for the first group the outcome is framed in terms of lives saved, while for the second group the problem has been framed in terms of lives lost. Tversky and Kahneman conclude that people tend to be risk averse when it comes to achieving

gains and risk taking when it comes to avoiding losses. Therefore, framing the problem in terms of either gains or losses influences how agents choose. Thus, nudges can influence behaviour by presenting information in a certain way, without changing the information. This effect is particularly prominent in digital environments, where digital nudges use personalization to create especially potent framing.

5.3 How digital nudges undermine information

Digital nudges undermine the information agents consume. Unlike for the other two limitations, cognitive abilities and willpower, there is virtually no existing literature on the relationship between (digital) nudges and information. We argue that digital nudges undermine the quality of information that agents consume by 1) causing an overload of information; and 2) causing the consumption of systematically biased information.

5.3.1 Overload of information

In standard economic literature, it is usually assumed that information is valuable, at least to the extent in which it aids decision making (Golman et al., 2017). Even if information has no clear benefit to decision making, information is assumed to be at least of neutral value as it can be ignored at no cost. It follows that "more information is always better" (Gigerenzer & Brighton, 2009, p. 110) and that "valid information should never be actively avoided" (Golman et al., 2017, p. 96).

However, humans are boundedly rational, they possess limited cognitive abilities and cannot process unlimited amounts of information. Simon notes that bounded rationality can be looked at from two perspectives. On the one hand, it could be looked at from the perspective of the environment. From this perspective, one might state that the environment provides too much

information. As Simon argues we live in “an information rich world” (Simon, 1971, p. 40). Reversely, bounded rationality could be looked at from the perspective of the agent’s capabilities. From this perspective, one might state that the agent’s capabilities to consume information are limited relative to the scarce resource which information consumes, namely attention. According to Simon an information rich world is equivalent to an attention scarce world.

As boundedly rational agents cannot assess all information provided by their environment, they need to divide their limited attention by filtering the different pieces of available information. Therefore, boundedly rational agents make decisions based on imperfect information. Some authors have concluded that agents working under conditions of imperfect information are faced with an optimisation under constraints problem (Arrow, 2004). As agents cannot process all information, they make decisions about which information to spend their attention on and hence which information to incorporate into their decision-making. As time and attention on the one hand and information and cognitive ease on the other hand are valuable, agents are faced with a trade-off between the value of spending additional time and attention and the value of an additional piece of information to improve decision-making. This trade-off can be referred to as the accuracy-effort trade-off (Wheeler, 2020).

Due to the accuracy-effort trade-off, it is rational to dismiss or ignore substantial amounts of information, to preserve one's attention span and save time. It follows that it is essential to consciously organize the environment in such a way that provides an optimal amount of information. Despite this need for optimal information environments, recent technological developments have moved us in the opposite direction.

Information technology plays a key role in increasing the amount of information agents consume. Firstly, the amount of data has grown explosively over the last years due to the development of information technologies (Taylor, 2024). Measurement, storage, computing, and analysis have all become both

cheaper and better at an exponential pace (Vopson, 2020). Secondly, agents spend more time in digital environments, which are informationally dense environments. Thirdly, the growth of the amount of data has made agents more reliant on the filtering capacities of IT. Algorithms ultimately determine which information becomes visible to agents as there is too much information for humans to filter all of it themselves. As a result, we have become more dependent on filtering algorithms. Fourthly, the same is true of digital nudges. Digital nudges highlight, accentuate and organize information, in short: they primarily determine how users navigate informational environments. With the growth of the amount of information and the corresponding relative decline in attention, dependence on digital nudges has increased in the same way as dependence on filtering algorithms.

For the first two developments it is clear how they contribute to an increase in information consumption. However, the increase in dependence on filtering algorithms and digital nudges need not increase information consumption. In fact, both filtering algorithms and digital nudges could be used to diminish information consumption (Okeke et al., 2018).

Yet in practice, the opposite occurs. As we have seen in a previous chapter, the dynamics of the digital economy create economic pressure on big tech corporations to capture as much user attention as possible. The more time users spend on a platform, the more advertisements the platform can show, and the more user data are generated to understand users better. As a result, the firms are directly incentivised to maximise the time users spend on their platforms.

Digital nudges are employed to present information in such a way that users keep consuming more. Even though digital nudges do not increase the amount of data that exists, they do increase the amount of information that people consume. Digital nudges can be used to attract attention, maintain attention and to highlight certain aspects of a design or of information. Furthermore, digital nudges are used to create feelings of wanting more by

delivering information in chunks via an intermittent reinforcement schedule. This creates addictive tendencies in users, craving more and newer information.

The cumulative effect of these design strategies is profound. In this way, digital nudges ensure that informational environments are constantly steering towards the consumption of more information rather than less, leading to the excessive consumption of information, which can be referred to as "information overload", which he defines as "a condition in which an agent has—or is exposed to, or is provided with—too much information, and suffers negative consequences as a result (experiences distress, [...] is unable to make a decision or to stay informed on a topic)" (Levy, 2008, p. 497). Information overload creates adverse effects for users as their environments prey on dysfunctional instincts. We can make a parallel with food. For our ancestors, an almost unlimited craving of sugar was helpful as the environment they lived in was tough a food was scarce. Starvation was a serious concern. Anytime they were presented with the chance to fill themselves up, they needed to seize it. For modern people, insatiability is a liability as the environment has changed and offers so much sugar that overconsumption is a bigger concern than starvation and the excessive intake of sugar is making us sick. Likewise, information overload causes adverse effects. These adverse effects include first, a reduction of cognitive abilities such as concentration, creativity and even intelligence, which we will analyse in chapter 6.

Second, information overload may cause the short-term effect of functioning poorly. When users are overwhelmed, it becomes difficult to properly execute simple tasks at hand. Feelings of being overwhelmed also reduce willpower, which we discuss in chapter 7.

Third, being overloaded or overwhelmed is an unpleasant sensation and is associated with stress and even anxiety (Levy, 2008).

Fourth, information overload increases the effectiveness of digital nudges as it forces people to fall back onto their shallow thinking and their biases (Levy, 2008). We revise this argument in more depth in chapter 6.

Beyond these immediate cognitive effects, there are broader informational consequences. Fifth, information overload causes users to be less informed as they cannot filter out the relevant information. Moreover, they may become less informed as they actively avoid information to prevent overload. Moreover, when people are confronted with many alternatives to choose from, they tend to perform worse due to the effort the decision requires. This phenomenon is referred to as "choice overload" or "overchoice" (Gourville & Soman, 2005). In general, "the choice process can be impeded by a glut as well as by a paucity of information" (McPherson, 1982, p. 266). This is especially important in the light of preference formation. To develop sound preferences, it is essential to understand one's options clearly. If there are too many options, choosers may not understand them, leading them astray.

Sixth, we have argued that an excess of information is equivalent to a dearth of attention. This dearth of attention causes people to respond to short-term demands constantly and limits the time people have for reflection on their values and actions. As David Levy argues: "when individuals, groups, or even entire societies become so enmeshed and overwhelmed by the amount of information and the accelerating pace of life [...] they fail to adequately develop the habits and character traits from which ethical action springs" (Levy, 2008, pp 511-512). The important habits and character traits that suffer from a dearth of attention according to Levy include patience, forgiveness, and empathy.

For proficient preference shaping, time and attention for reflection are particularly important because they allow agents to evaluate the consequences of their choices over longer periods of time. Information overload restraints attention and thereby reflection, leading to underdeveloped preferences and hence a loss of welfare.

Seventh and finally, the dearth of attention "may turn our attention from what is important to what is unimportant. We cannot afford to attend to information simply because it is there" (Simon, 1978, p. 13). This final point serves as a critical bridge to our next concern. In the next section, we turn to the kinds of information that thrive due to digital nudges, and we will argue that digital nudges overwhelm people particularly with unimportant and even harmful information.

5.3.2 Biased information

Digital nudges harm welfare as they cause the consumption of systematically biased information. In the digital economy, there is a competition for user attention. Information technology uses filtering algorithms and digital nudges to select information with the intent of capturing attention. We have argued in chapter 3 that such ordering is a form of digital nudging if it is performed with the goal of steering user behaviour, typically aligned with the economic incentives of digital designers to maximize engagement and attention. As a result, the kinds of information and presentation-forms which capture attention most successfully are favoured, causing the informational environment to be biased in a systematic and predictable manner.

Hills analyses information from an evolutionary perspective, arguing that information answers to the "forces of cognitive selection" (Hills, 2019, p. 323). According to him, "information that is more likely to be searched for, attended to, comprehended, encoded, and reproduced" is favoured and hence more likely to be consumed and communicated. In addition, he argues the increase in the amount of information, which has taken place as we analysed in the previous section, affects "the evolution of information" through two mechanisms (Hills, 2019, p. 324). First, the increase in information increases the competition for attention. As there is more information which competes for the same amount of attention, the prowess of selection mechanisms rises. Second, as information technologies have reduced the cost and time required for the creation of

information, information travels faster between people. This also strengthens the cognitive selection mechanisms since the velocity with which information is created and communicated enforces greater adaption to the selection mechanisms. As Hills formulated it: "[t]he more rapidly that people can access, select, and reproduce preferred information, the more readily will that information reflect the cognitive biases of its users" (Hills, 2019, p. 324).

Building on Hills' evolutionary perspective, we argue that in a digital world, information technologies–consisting of filtering algorithms and digital nudges–are specifically designed to evolve toward whatever captures most attention. This means that information technologies come to mirror the cognitive selection mechanisms of their users. As argued in chapter 3, these technologies work together to filter and highlight information to assist users in navigating the enormous amount of it and so prevent information overload. However, the mechanisms by which this filtering occurs remain largely hidden from users. This algorithmic invisibility prevents users from adequately assessing how the information presented to them was selected, further exacerbating biases.

The issue of biased information has been extensively discussed in the literature on the ethics of search engines (Tavani & Zimmer, 2025). More specifically, it has been argued that search engines are not neutral and favour certain values over others. According to Brey these values remain "hidden" and must be made apparent with "disclosive analysis" by (computer) ethicists (Brey, 2000, p. 13). Moreover, search engines are accused of systematically favouring certain (kinds of) websites over others and of failing to use objective criteria in the selection and ordering of search results (Tavani & Zimmer, 2025). Certain kinds of information are favoured, yet how and why remains opaque.

This opacity operates across digital platforms. It has been argued that personalization and recommender algorithms serve users only content which confirms the user's political views and shield them from opposing views, thereby undermining civic discourse and fuelling polarization (Pariser, 2011). However,

various studies have been conducted to assess whether filter bubbles exist, and the evidence is mixed at best (Haim et al., 2018; Nechushtai & Lewis, 2019; Zuiderveen et al., 2016). In any case, information on social media must answer to the dynamics of digital markets, which dictate that information is only selected based on its probability that it will keep the user on the platform.

Likewise in search engine optimization (SEO), website owners try to optimize their content in a way that makes it more likely to be favoured by search engines. While Google's selection criteria are secret (algorithmic invisibility) and subject to continuous change, there are numerous commonly accepted rules that website owners typically adhere to, including website structure, content length, and use of specific words. Importantly, the Google algorithm favours websites with high clickthrough rates (when a user clicks any link on the web page) and dwell times (the amount of time a user spends on the page). As a high ranking in Google is vital to being found online and thus of exceptional value, many people adhere to Google's rules and aim to optimize these metrics. The massive attempt to game Google's algorithm has thereby shaped the structure of the web and created a bias in how information is presented, and which information is visible.

These examples of algorithmic influence across search engines and social media platforms illustrate a broader pattern: if users have a bias for consuming or sharing certain kinds of information, digital nudges systematically amplify this cognitive bias. In other words, digital nudges will disproportionately push the kinds of information that users are already predisposed to consuming and sharing.

In analysing these systematic patterns, Hills identifies four categories of information which enjoy preferential selection and are hence consumed disproportionately, namely information that is 1) belief-consistent; 2) social; 3) negative; or 4) predictive. In the following sections, we extend his argument by discussing how the preferential selection of these four categories of information

are affected by digital nudges. In addition, we extend on Hills by arguing for a fifth category: empty information.

5.3.2.1 Belief consistent information

First, belief-consistent information is favoured by digital nudges as users have a tendency "to seek out, encode in memory, and reproduce information consistent with prior beliefs" (Hills, 2019, p. 324). People naturally seek out information that is consistent with existing beliefs and possibly avoid messages that oppose their beliefs. This tendency becomes stronger as people are confronted with larger quantities of information (Fischer et *al.,* 2008), as they are as analysed in the previous section. Algorithms and digital nudges play into the tendency to seek out confirmative information, reinforcing it even further. As a result, people live in informational environments that are excessively confirmative with an equivalent lack of belief-inconsistent information.

5.3.2.2 Social information

Digital nudges favour social information (Hills, 2019). Social information is all information pertaining to conduct of others. There is a fundamental "human need to monitor and be monitored by others" (Veissière & Stendel, 2018, p. 1). So, people have a tendency to consume information regarding what others do. Digital nudges use social information to attract attention. For example, headlines of news stories are often formulated in terms of personal narrative. Headlines don't read 'a storm occurred' but 'John, aged 42, was caught in a storm'. Another example of a digital context that nudges towards social information is social media platforms. As social media corporations know that people are likely to respond to social information, they digitally nudge people to share information concerning themselves, as they know this will increase the attachment of both the consumers of this information, as well as the providers of the information

who receives likes (Facebook), hearts (Instagram), or some other digital token of social appreciation.

The highly social nature of social media is associated with various types of harm. There is evidence that people, especially young people, crave social appreciation to such a great extent that they develop smartphone addiction (Veissière & Stendel, 2018). In addition, the social nature of a large amount of information invites people to compare themselves to others. There is research which connects Facebook-use to social comparison, envy, and depression (Appel et al., 2016). While it stands to reason that excessive social comparison is detrimental to one's mental health, social media incorporate several features that make social comparison particularly detrimental to a good informational position. First, social media show a highly biased subset of people. Second, social media allows to portray a false image of oneself. Third, the group one compares oneself to is too large. Comparing oneself to a larger reference group leads to a less fulfilling life as there is always someone on YouTube better than you.

Moreover, people not only consume social information but also make decisions based on the behaviour and opinions of others (Cialdini, 1984). This phenomenon is referred to as conformity bias, social proof, or informational social influence. To illustrate the power of social influence, Goldstein *et al.* refer to research which investigated which messages stimulate hotel visitors to reuse their towel. Of the guests who were presented with the message "HELP SAVE THE ENVIRONMENT. You can show your respect for nature and help save the environment by reusing your towels during your stay", 35% reused their towel (Goldstein *et al.*, 2008, p. 473). However, 44% of the guests who were presented with the message "JOIN YOUR FELLOW GUESTS IN HELPING TO SAVE THE ENVIRONMENT. Almost 75% of guests who are asked to participate in our new resource savings program do help by using their towels more than once. You can join your fellow guests in this program to help save the environment by reusing your towels during your stay" (Goldstein *et al.*, 2008, p. 474).

Pointing to the behaviour of others is a "social nudge" and is considered one of the most potent forms of nudging (Thaler and Sunstein, 2008, p. 55). According to Thaler and Sunstein, there are two main reasons why people exhibit conformity and like to follow the herd. First, the behaviour of others is interpreted as informational evidence of the best alternative (Thaler and Sunstein, 2008). Second, the behaviour of others exerts peer pressure on people, as people have a deep longing to be accepted and appreciated (Thaler and Sunstein, 2008).

Digital nudges favouring social information can lead to herding behaviour and disrupt the wisdom of the crowd (Lorenz *et al.*, 2011). This is true of all social information and all social nudges. Nevertheless, digital social nudges aggravate the problem due to the three reasons listed above.

5.3.2.3 Negative information

Third, digital nudges favour negative information (Hills, 2019). As a remnant from their evolutionary history, humans are highly sensitive to disadvantages, risks, and negative information (Hills, 2019). This amounts to a "negativity bias", defined as a tendency to "attend to, learn from, and use negative information far more than positive information" (Vaish, et al., 2008, p. 383). As a result, digital nudges favour negative information to capture and maintain the attention of users.

Moreover, digital nudges are used to accentuate risks with the aim of motivating behaviour. Tversky and Kahneman (1991) describe how choices are impacted by loss aversion and reference dependence. Loss aversion (or risk aversion) is a bias describing that people weigh losses heavier than corresponding gains, a widely found and well documented phenomenon (Tversky & Kahneman, 1991). Reference dependence is the phenomenon that people evaluate choices relative to a reference point when determining their

preferences. Both phenomena are used to digitally nudged people. With respect to loss aversion, digital nudges can be used to frame choices in terms of loss rather than gains. For example, many marketing slogans use phrases such as "don't miss out" or "limited offer" (using scarcity bias). Another example with digital products is to offer it free for a limited time. Once consumers have (access to) a digital product, they don't enjoy losing it (using the endowment effect). With respect to reference dependence, digital nudges can be used to set or alter the reference point people use when making a decision to modify their preferences in this way. For example, many online vendors show the previous price of a product as a point of reference, which functions as an anchor. Then, they show the discounted price which seems cheap by comparison. By setting the reference point, the vendor is in this way able to digitally nudge the consumer and make them perceive the product offer favourably. Framing the reference point can be used to make someone perceive something as a loss.

Both the tendency of digital nudges to supply an excess of negative information and the use of loss aversion and reference dependence to make users perceive more losses lead to an informational environment in which people perceive losses and risks as bigger than they are. Overexposure to, or overestimating the importance of, negative information leads to risk averse behaviour and to "preference reversals" or context-dependence of preferences (Tversky & Kahneman, 1991).

5.3.2.4 Predictive information

Fourth, digital nudges favour predictive information (Hills, 2019). Humans have a "bias for pattern identification" (Hills, 2019, p. 326) which leads to apophenia, the "tendency to perceive a connection or meaningful pattern between unrelated or random things" (Merriam-Webster, n.d.). As a result, the likelihood of the existence of a positive relationship between two variables is

overestimated. This problem is exacerbated by the explosive growth in the amount data because an increase in the amount of data leads to an increase in the potential for spurious and meaningless correlations.

Digital nudges promote predictive information because it contributes to their goal, capturing the attention of users. As such, digital nudges favour predictive information over other information, leading to an incentive to promote falsely positive information. In other words, to stimulate misinformation. This becomes especially apparent when reviewing YouTube's recommendation algorithm which is infamous for recommending misinformation and sending people down the "rabbit-hole" as those who have already watched misinformation, were more likely to be recommended further video's containing falsehoods (Hussein *et al.*, 2020). The critique on YouTube is perhaps best formulated by ex-employee Guillaume Chaslot, who witnessed first-hand how YouTube's recommendation algorithms started "pushing users toward conspiracy videos" (Popken, 2018). He attested the algorithm "figured out that the best way to get people to spend more time on YouTube was to show them videos light on facts but rife with wild speculation" (Popken, 2018). Despite efforts to curb misinformation, the platform has been criticized for promoting falsehoods and conspiracy theories.

5.3.2.5 Empty information

In our view, the four categories of information as mentioned by Hills (2019) are incomplete. We argue digital nudges favour a fifth type of information, which we will refer to as empty information. It becomes clear what empty information is when analysing the presentation of digital information, whether it is on YouTube, Netflix, Google search, Facebook, Instagram, TikTok, Amazon, or another digital platform. Information found here is presented in a way that is best described as effortless, palatable, fast, superficial, and shallow,

often baiting curiosity without fulfilling it. It is designed to be quickly and easily consumed, often prioritizing engagement and viewer retention over the provision of substantive information. As such, empty information is the informational equivalent of empty calories, as the calories provided by fast food are called. Fast-food seems attractive and offers the promise of instant gratification. Yet, it lacks content, is unnourishing, and provides an unfulfilled, empty feeling. It is a fleeting pleasure, leaving one to crave more.

An example of empty information favoured by digital nudges are click-bait titles, easy-read formats, and blogposts written as numbered lists. Clickbait titles over-promise the content of the article and leave the reader often unfulfilled and disappointed. Moreover, easy-read formats divide information into tiny bite-sized chunks to ask as little commitment from the reader as possible. As a longer piece of text appears daunting, online media use digital nudges to make the text feel manageable. For example, many texts now have the reading time stated at the top, a progress bar to convey to readers how far they are along in the text, or the division of text into ultrashort paragraphs, sometimes consisting of no more than a single sentence throughout the entire text. In addition, when using a search engine for information, one frequently finds pages consisting of a numbered list, such as '6 things you must know', '7 tips against problem x', or '8 ways to improve x'. These pages are designed to be easily skimmable, and to quickly provide small amounts of shallow information.

Another good example of empty information is the progressively shorter, faster and more easily accessible video formats found on social media. TikTok videos are typically between 15 and 60 seconds and are presented in an endless timeline which is seamless to a degree that scrolling through it becomes both intuitive and automatic grace to the development of muscle memory in the user's thumb. This format is likewise found on YouTube (YouTube shorts), Instagram and other social media. And the endless stream of video's is literally infinite. Instagram used to have an end to their feed but changed with the goal of

“nudging users to spend more time on the app so Instagram can serve them more ads [...]. Now that the infinite scroll is actually infinite, it will never end—and neither will the ads” (Gagne, 2020, para. 10).

These infinite scrolls are so captivating that one easily spends hours on end watching and scrolling. The success of this format can be attributed to several features and psychological mechanisms encapsulated in these platforms. First, the machine learning algorithms utilised by the platform to select the best – that is most attention capturing – videos are highly proficient. YouTube’s recommendation algorithm has been described as a “supercomputer playing chess against your mind to get you to keep watching” (Popken, 2018, para. 1).

Second, infinite scrolls leverage the power of intermittent reinforcement, establishing an eager want in people by providing rewards in unpredictable patterns, causing users to anticipate rewards. As described in a previous chapter, this is associated with addictive behaviour. They institute the hope in users of finding better – that is more pleasure – content later on in the scroll, induce them to continue scrolling.

Third, this format erases the effort of committing. Many people have the experience that choosing, starting, and committing to watching a movie, which typically lasts between 90 and 150 minutes, costs effort, while watching short clips on social media for the corresponding duration of a movie is effortless.

Fourth, the instant nature of the scrolling ensures that immediately when users are bored, irritated, or otherwise negatively affected, they can change the content with the swipe of a thumb, ensuring they do not need to cope with the slightest of negative feelings for even a moment. As a result, they may come to associate slight, almost unconscious negative feelings with the action of swiping their thumbs to summon the next video (Eyal, 2014). In fact, they may even come to associate negative feelings like boredom, anxiety, or indecisiveness in general with swiping through an endless scroll, even when not using their phones (Eyal, 2014).

The combination of the above ensures the user enters a flow state. This is a mental state defined by intense "concentration on the task at hand" (Csikszentmihalyi, 2004, p. 136), "loss of self-consciousness" (p. 146), "the transformation of time" (p. 153), "the merging of action and awareness" (p. 126), and a (possibly false) "sense of control" (p. 141). When in a state of flow, people have complete focus on the task at hand, to the extent that "one is able to forget all the unpleasant aspects of life" and about oneself (p. 136). The flow state is fully consuming attention to the extent that literally no other information than the task at hand enters the mind, including any minor sign of boredom, anxiety, indecisiveness, sadness or any other negative emotion.

This total immersion into a mental activity that shuts out all forms of inner disorder and discomfort explains why the flow state is potentially addictive (Csikszentmihalyi, 2004). Natasha Dow Schüll (2012) connects the flow state to problematic gambling behaviour, especially to slot-machine games. While Csikszentmihalyi grosso modo has a positive evolution of the flow state, Schüll explains there is a harmful variant of flow, the one gamblers experience which is "depleting, entrapping, and associated with a loss of autonomy" (p. 167). She argues that the same principles which were utilized by the designers of slot-machines to make them maximally addictive are used by social media for the same purpose (Harris & Raskin, 2019). Social media, especially those with infinite scrolls, have indeed proven to be addictive (Sun & Zhang, 2021; Hou *et al.*, 2019; Cheng *et al.*, 2021), leading to the consumption of large amounts of empty information.

Empty information can obviously lead to reduced welfare as addiction and dependence cause substantial harm and are associated with depression (Twenge *et al.*, 2018a; Twenge *et al.*, 2018b), although it is not clear whether the social media addiction causes the depression or vice versa (Longstreet & Brooks, 2017). Yet, we aim to focus on the less obvious welfare effects of shallow information, including for those who do not develop depressive symptoms.

Our present discussion of empty information illustrates that the type of information that captures attention is not necessarily valuable. In fact, the excess of empty information due to digital nudges captures an excessive portion of user attention to the extent that it leads to a substantial decrease in the consumption of substantive and rich information, which requires effort. In addition, people who consume an excess of empty information may come to prefer only the shallow, coming to prefer distraction rather than true engagement. As such, the selective preferential treatment of empty information by digital nudges can cause an unintentionally acquired preference for the empty and shallow.

5.4 Digital nudges proficient preference shaping

In Chapter 4, we established that welfare depends on the satisfaction of preferences that have been proficiently shaped through experimentation and reflection. This process rationality approach recognizes that preferences are not fixed but evolve through engagement with the world, like a blade sharpened against a stone. For this sharpening process to occur, the informational environment must function properly.

In this chapter, we demonstrated how digital nudges systematically distort the informational environment through overload and bias. Building on this analysis, we now examine more deeply how these informational distortions specifically undermine proficient preference shaping. The negative impact on information manifests in two aspects: distortion of information regarding available alternatives and disrupted feedback on choices.

First, agents need information regarding the alternatives they can choose from. Second, agents need feedback, receiving information to understand the results of their choices. Yet, digital nudges distort both components, undermining the process rationality that is essential for proficient preference shaping.

5.4.1 Distortion of information regarding available alternatives

An understanding of one's alternatives is hampered by the biased informational environment agents find themselves in as certain types of information are systematically favoured. Agents do not engage with a sufficiently large set of alternatives as their environment favours belief-consistent, social, negative, predictive and empty information. Simply put, people lack diversity as they are not confronted sufficiently with belief-inconsistent, positive, and rich information. Personalization exacerbates this mechanism, as it algorithmically filters potential alternatives for agents, ensuring they may never encounter them. Consequentially, the higher, richer and more effortful alternatives are less understood, which negatively affects one preference shaping. When agents lack meaningful exposure to richer and more demanding forms of life, their capacity to form, revise, and stabilize higher-order preferences is weakened. Consequently, the process of preference shaping is undermined.

As Mill states: people "addict themselves to inferior pleasures, not because they deliberately prefer them, but because they are either the only ones to which they have access, or the only ones which they are any longer capable of enjoying" (Mill, 2001/1863, p. 13). This is exactly what happens when digital nudges distort the informational environment by overflooding it with an overload of information, which is biased towards the empty, shallow and self-confirming. One comes to prefer the shallow and empty because it is easy, and after continuous reinforcement becomes habitual. The shaping of preferences towards more diverse and perhaps more effortful pleasures becomes suppressed and the higher more effortful preferences suffer.

In addition, information overload increases the costs of finding correct and relevant information as the bigger the amount of information a user is confronted with, the more difficult it becomes to select. Moreover, there is

evidence that information overload can lead to information avoidance or information blindness (Golman et al., 2017), limiting agent understanding of their alternatives. In the context of markets, this can cause consumers to be blind or unwilling to perceive warnings and terms and conditions (Calo, 2014).

5.4.2 Distortion of feedback mechanisms

The second way digital nudges distort the informational environment of agents and thereby proficient preference shaping is through distortion of feedback mechanisms. When agents perform experiments in living explore what they might prefer, they might find activities they delight in and others which they dismay. In other words, for effective experimentation and proficient preference shaping they must be aware of their affective responses to the experiments they conduct. Fabian and Dold call the emotional responses "affective feedback" (2022, p. 9). These authors also stress that efficient feedback loops require attention and reflection. Understanding how one's choices have affected one's utility, and to evaluate one's life as a partial result of one's own choices is a daunting task requiring skill, attention and conscious reflection. Both attention and effortful reflection are supressed by the continuous barrage of digital nudges towards more information consumption. As we have described, intermittent reinforcement and clever digital design have been constructed to link the slightest feelings of boredom, anxiety and frustration to the need to scroll through social media, completely overshadowing all moments of quiet reflection.

Finally, another mechanism harming the effective feedback mechanism is the progressively shorter cycle one expects the rewards of their actions in. The quick rewards offered by complying with digital nudges in the form of dopamine hits are practically instantaneous. The instant gratification of digital nudges

makes them an unfair opponent for other, perhaps more intrinsically enjoyable activities, of which the feedback loop is much longer.

Ultimately, as agents do not properly understand their alternatives and do not have functioning feedback mechanisms to acquire a better understanding of the choices they make, their self-knowledge suffers. In Chapter 4, we described how, through dealing with inconsistencies and incompleteness and by experimenting into the unknown, agents learn "to perceive the interconnectedness of [...] interests so as to pursue multiple ends harmoniously and coherently" (Fuchs, 2001, p. 237). Agents shape their preferences and gain self-knowledge of what they prefer through reflecting on their actions. Agents lacking such an understanding or acting against their own preferences can be described as “fragmented selves” (Bovens, 2009, p. 8). As digital nudges ensure information is biased, and affective feedback is distorted, agents are much more likely to become such fragmented selves, and their welfare suffers as a result.

5.5 Recommendations

To mitigate the harmful effects of digital nudges on information environments and thereby on proficient preference shaping and welfare, we propose the following recommendations targeted at high schools and educational policymakers. These recommendations aim to equip students with the knowledge and skills needed to navigate digital information environments that are increasingly shaped by digital nudges.

5.5.1 Develop curricula on understanding digital information environment dynamics

Educators should develop curricula that help students understand the dynamics that shape digital information environments. Media literacy modules should teach students how to recognise the incentive structures behind digital platforms they regularly use, helping them understand why certain content is

promoted over others. Through these modules, students can gain insight into the commercial motivations that drive content curation and learn to identify when their attention is being deliberately captured.

Teachers should guide classroom discussions about the trade-off between information and attention, emphasising that while information is abundant, attention is scarce and valuable. These discussions can include practical exercises where students track their digital consumption and analyse how their attention is directed. Such exercises help students recognise patterns in their own usage and develop greater awareness of how their attention is being allocated across different digital platforms.

Educational programmes should include analysis of how digital platforms filter and organise information, with students examining how their personal feeds are curated by algorithms and how this affects their informational position. By comparing different users' feeds on the same platform, students can observe firsthand how algorithmic filtering creates individualised information environments and learn to consider the implications of these differences.

Schools should incorporate reflective exercises where students consider how their informational environment affects their preferences and self-understanding, encouraging them to question whether their digital habits are conducive to their welfare. These exercises might involve periodic digital journals where students document changes in their interests and preferences in relation to their digital consumption habits.

Media courses should provide clear, age-appropriate explanations of digital nudges, how they shape informational environments, and their potential adverse effects, using real-world examples relevant to teenagers. By examining case studies of popular platforms, students can learn to identify specific nudge techniques and discuss their psychological impact.

5.5.2 Teach mindful information consumption techniques to prevent overload

According to Bawden and Robinson, information overload can "best be overcome by seeking a mindful balance in consuming information, and in finding understanding" (2020, p. 1). To help students manage information overload exacerbated by digital nudges, schools should develop practical workshops on mindful information consumption, teaching students the techniques to consciously limit their information intake. These workshops can introduce students to strategies such as time-blocking, content filtering, and selective consumption, providing them with concrete tools to manage information flow.

Digital citizenship curricula should specifically identify high-risk applications, such as those without friction, with infinite supply, or with intermittent reinforcement, and discuss their impact on attention and information processing. Through analysis of app design features, students can learn to recognise applications that are particularly likely to contribute to information overload and develop strategies for moderating their use.

Teachers should demonstrate how to be comfortable during brief waiting periods without automatically turning to smartphones for stimulation. By normalising momentary inactivity and even boredom, educators can help students develop tolerance for reduced information input, which is essential for combating the effects of digital nudges that constantly push for engagement.

5.5.3 Guide students in seeking diverse and deep information sources

To counter the biasing effects of digital nudges on information exposure, library and information science classes should teach students how to actively seek diverse information sources, helping them develop research skills that extend beyond algorithm-curated content. These classes can introduce students

to academic databases, specialised information sources, and alternative search engines that may provide different perspectives from mainstream platforms. It should also emphasize the importance of experimenting in life, in line with the experiments in living discussed in chapter 4.

Critical thinking courses should examine how digital nudges tend to promote certain types of shallow information, with students learning to identify when their consumption patterns are being narrowed by algorithms. Through content analysis exercises, students can compare depth and diversity across different information sources and recognise when they are being nudged toward superficial engagement.

Schools should design projects that require students to deeply engage with topics of personal interest, helping them develop the skill of dedicating sustained attention to valuable information rather than passively consuming what digital nudges promote. These projects can be structured to encourage progressive depth of engagement, counteracting the tendency of digital environments to promote shallow browsing across multiple topics.

5.5.4 Train students to customize digital tools and manage notifications

To help students implement practical changes through digital tools, computer science or digital literacy courses should include components where students audit their own technology use, reflecting on their reasons for using specific digital technologies. Through guided self-assessment, students can develop greater awareness of their technology choices and consider whether these align with their values and goals.

Educational programmes should teach students how to assess digital tools based on their values, providing frameworks for assessing the benefits and drawbacks of different applications, particularly regarding digital nudges. By

developing evaluation criteria related to attention demands, information quality, and customisation options, students can make more informed choices about which digital tools to incorporate into their lives.

Schools should offer practical workshops on customising digital tools, teaching students how to adjust settings to create healthier information environments. These hands-on sessions can guide students through the process of modifying default settings on common platforms to reduce distractions, limit notifications, and control information flow.

Digital literacy curricula should specifically address notification management, with students learning to mindfully limit notifications to reduce information overload and addictive patterns. By examining the psychological effects of different notification types and experimenting with various notification settings, students can develop personalised strategies for managing digital interruptions.

Schools should introduce students to third-party tools that allow greater customisation of popular platforms, teaching them how to use browser extensions and other utilities to modify their digital environments when direct customisation options are limited. These lessons can include practical demonstrations of tools that block recommended content, remove addictive features, or provide alternative interfaces for common platforms, like timeline eradicators for YouTube.

Schools should involve parents in digital literacy education, providing guidance on implementing consistent digital practices at home that reinforce what students learn in school. Through parent workshops, information sessions, and take-home resources, schools can extend their influence beyond the classroom and help create more coherent digital environments for students.

By implementing these recommendations within secondary education, schools can equip students with the skills needed to navigate information

environments increasingly shaped by digital nudges, fostering more proficient preference shaping and ultimately enhancing student welfare.

5.6 Bibliography

Appel, H., Gerlach, A. L., & Crusius, J. (2016). The interplay between Facebook use, social comparison, envy, and depression. *Current Opinion in Psychology*, 9, 44–49. https://doi.org/10.1016/j.copsyc.2015.10.006

Arrow, Kenneth. (2004). Is Bounded Rationality Unboundedly Rational? Some Ruminations. In M. Augier & J. G. March (Eds.), *Models of a Man: Essays in Memory of Herbert A. Simon* (pp. 47–55). MIT Press.

Bawden, D., & Robinson, L. (2020). Information overload: An overview. In *Oxford encyclopedia of political decision making*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.1360

Bovens, L. (2009). The ethics of nudge. In Preference change: *Approaches from philosophy, economics and psychology* (pp. 207-219). Dordrecht: Springer Netherlands.

Brey, P. (2010). Philosophy of technology after the empirical turn. *Techné: Research in Philosophy and Technology*, 14(1), 36–48.

Calo, R. (2014). Code, nudge or notice. Iowa Law review, 99, 773.

Cheng, C., Lau, Y., Chan, L., & Luk, J. W. (2021). Prevalence of social media addiction across 32 nations: Meta-analysis with subgroup analysis of classification schemes and cultural values. *Addictive Behaviors*, *117*, 106845.

Cialdini, R. B. (1984). *Influence: The psychology of persuasion*. HarperCollins.

Csikszentmihalyi, M. (2004). Good business: Leadership, flow, and the making of meaning. Penguin.

Eyal, N. (2014). Hooked: How to build habit-forming products. Penguin.

Fabian, M., & Dold, M. (2022). Agentic preferences: A foundation for nudging when preferences are endogenous. *Behavioural Public Policy*, 1–21. https://doi.org/10.1017/bpp.2022.17

Fischer, P., Schulz-Hardt, S., & Frey, D. (2008). Selective exposure and information quantity: How different information quantities moderate decision makers' preference for consistent and inconsistent information. *Journal of Personality and Social Psychology*, 94(2), 231-244.

Fuchs, A. E. (2001). Autonomy, slavery, and Mill's critique of paternalism. *Ethical Theory and Moral Practice*, 4, 231–251.

Floridi, L. (2010). *Information: A very short introduction*. Oxford University Press.

Gagne, Y. (2020, August 19). Instagram's new suggestion feature ensures your feed will never end. Fast Company. https://www.fastcompany.com/90541401/instagrams-new-suggestion-feature-ensures-your-feed-will-never-end

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, 1(1), 107-143.

Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of economic literature*, 55(1), 96–135.

Goldstein, N., Cialdini, R., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, 35(3), 472-482.

Goldstein, D., Johnson, E., Herrmann, A., & Heitmann, M. (2008). Nudge your customers toward better choices. Retrieved from https://hbr.org/2008/12/nudge-your-customers-toward-better-choices

Gourville, J. T., & Soman, D. (2005). Overchoice and Assortment Type: When and Why Variety Backfires. *Marketing Science*, 24(3), 382–395. https://doi.org/10.1287/mksc.1040.0109

Haim, M., Graefe, A., & Brosius, H.-B. (2018). Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digital Journalism*, *6*(3), 330–343.

Harris, T., & Raskin, A. (Hosts). (2019, February 4). What happened in Vegas (No. 1) [Audio podcast episode; guest Natasha Dow Schüll]. In *Your Undivided Attention*. Center for Humane Technology. https://www.humanetech.com/podcast/1-what-happened-in-vegas

Hills, T. T. (2019). The Dark Side of Information Proliferation. *Perspectives on Psychological Science*, *14*(3), 323–330. https://doi.org/10.1177/1745691618803647

Hou, Y., Xiong, D., Jiang, T., Song, L., & Wang, Q. (2019). Social media addiction: Its impact, mediation, and intervention. *Cyberpsychology*, 13(1).

Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1), 1–27.

Kahneman, D., & Tversky, A. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453–458.

Kahneman, D., & Tversky, A. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106(4), 1039–1061.

Levy, D.M. (2008). Information Overload. In *The Handbook of Information and Computer Ethics* (eds K.E. Himma and H.T. Tavani). https://doi.org/10.1002/9780470281819.ch20

Longstreet, P., & Brooks, S. (2017). Life satisfaction: A key to managing internet & social media addiction. *Technology in Society*, 50, 73-77.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22), 9020–9025.

Merriam-Webster. (n.d.). Apophenia. In Merriam-Webster.com dictionary. Retrieved September 19, 2023, from https://www.merriam-webster.com/dictionary/apophenia

Nechushtai, E., & Lewis, S. C. (2019). What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior*, *90*, 298–307.

Okeke, F., Sobolev, M., Dell, N., & Estrin, D. (2018). Good vibrations: Can a digital nudge reduce digital overload? Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services, 1–12.

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin UK.

Popken, B. (2018, February 7). As algorithms take over, YouTube's recommendations highlight a human problem. NBC News. https://www.nbcnews.com/tech/social-media/algorithms-take-over-youtube-s-recommendations-highlight-human-problem-n867596

Schüll, N. D. (2012). *Addiction by design: Machine gambling in Las Vegas*. Princeton University Press.

Sequoiah-Grayson, S. & Floridi, L. (2022). Semantic conceptions of information. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022 Edition). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2022/entries/information-semantic/

Simon, H. A. (1971). Designing organizations for an information-rich world. *Computers, Communications, and the Public Interest*, *72*, 37.

Simon, H. A. (1978). Rationality as process and as product of thought. *The American Economic Review*, *68*(2), 1–16.

Sun, Y., & Zhang, Y. (2021). A review of theories and models applied in studies of social media addiction and implications for future research. *Addictive Behaviors*, 114, 106699.

Tavani, H., & Zimmer, M. (2025). Search engines and ethics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2025 edition). Stanford University. https://plato.stanford.edu/archives/spr2025/entries/ethics-search/

Taylor, P. (2024). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025. Statista. https://www.statista.com/statistics/871513/worldwide-data-created

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Penguin.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453–458.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106(4), 1039–1061.

Twenge, J. M., Joiner, T. E., Martin, G., & Rogers, M. L. (2018a). Digital media may explain a substantial portion of the rise in depressive symptoms among adolescent girls: response to Daly. *Clinical psychological science*, 6(3), 296-297.

Twenge, J., Joiner, T., Rogers, M., & Martin, G. (2018b). Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time. *Clinical Psychological Science*, 6(1), 3–17.

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383–403. https://doi.org/10.1037/0033-2909.134.3.383

Veissière, S. P., & Stendel, M. (2018). Hypernatural monitoring: A social rehearsal account of smartphone addiction. *Frontiers in psychology*, 9, 331065.

Vopson, M. M. (2020). The information catastrophe. AIP Advances, 10(8), 085014. https://doi.org/10.1063/5.0019941

Wheeler, G. (2020). Bounded Rationality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/bounded-rationality/

Zuiderveen Borgesius, F. J., Trilling, D., Möller, J., Bodó, B., De Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles?. *Internet policy review*, 5(1), 1-16.

Chapter 6 – How digital nudges skew cognitive abilities

6.1 Introduction

Having examined information as the first of three fundamental limitations in the previous chapter, we now turn to the second limitation: how digital nudges skew cognitive abilities. We conceptualize cognitive abilities through the lens of dual process theory, relying mostly on Kahneman's (2011) work as it is central to nudge theory and Stanovich (2004; 2011) as he provides important nuances to the workings of system 2.

We then analyse how digital nudges impact cognitive abilities. We argue digital nudges fundamentally cause agents to overuse system 1 at the expense of system 2. We distinguish three welfare-harming mechanisms that arise because of this overreliance. First, it causes agents to disproportionately prioritize lower-order preferences, even when conflicting with higher-order preferences. Second, it causes agents to act for the wrong reasons as they are overly susceptible to biased thinking, hurting proficient preference shaping. Third, it causes the long-term deterioration of reflective and evaluative skills which are needed for shaping preferences as argued in chapter 4.

We conclude the chapter by providing recommendations on how teachers and educational policy makers can teach students to protect and improve cognitive abilities when confronted with digital nudges, with the goal of improving welfare through more proficient preference shaping.

6.2 Defining cognitive abilities based on dual process theory

Cognitive abilities are “mental processes involved in the acquisition of knowledge, manipulation of information, and reasoning. [They] include the

domains of perception, memory, learning, attention, decision making, and language" (Kiely, 2014, p. 975). Cognitive abilities have an obvious connection to welfare as they are not only needed to form and adjust true beliefs and to satisfy preferences but also to proficiently shape preferences, as explained in chapter 4. In this chapter, we assess how digital nudges affect cognitive abilities and thereby the proficient shaping of preferences and welfare.

Cognitive abilities also have an obvious connection to digital nudges since we defined digital nudges, based on Saghai (2013), as follows: A (digitally) nudges B when A makes it more likely that B will ϕ, primarily by triggering B's shallow cognitive processes, while A's influence preserves B's choice-set. So, digital nudges are defined by how they affect a particular type of cognitive processes, which Saghai refers to as "shallow" (2013, p. 489). Shallow cognitive processes here refer to one side of dual process theory. In the literature on nudging, shallow cognitive processes are commonly referred to as system 1 (type processes). Stating nudges target shallow cognitive processes can inversely be formulated as nudges avoiding robust cognitive processes, which are commonly referred to as system 2 (type processes). To achieve an understanding of how digital nudges affect welfare through their effect on cognitive abilities, we must first understand these two systems and how digital nudges relate to them.

6.2.1 System 1 thinking

Saghai describes system 1 processes as "shallow cognitive processes", which are "nondeliberative and incompletely deliberative processes that share three properties: (1) they are fast; (2) the 'cognitive miser' is inclined to rely on them because they consume few resources; (3) they yield responses that are not the result of full-blown deliberation" (Saghai, 2013, p. 489).

To experience the workings of system 1, answer the following questions: what is the opposite direction of left? What is 2 + 2? In both cases, the answer comes to mind immediately, automatically and effortlessly, which are features

of system 1 processing (Kahneman, 2011). As system 1 operates automatically and unconsciously, system 1 decisions are experienced as providing "no sense of voluntary control" (Kahneman, 2011, p. 21). They occur instantly without effort. Moreover, system 1 is associative (Stanovich, 2011) and intuitive (Evans & Stanovich, 2013). Finally, system 1 is highly contextualised. It works with the information provided by the direct context and is grounded in it. It is an experience-based form of decision-making.

System 1 is "heuristic processing" (Stanovich, 2011, p. 29). A heuristic is an efficient cognitive strategy that ignores part of the available information (Gigerenzer, 2009). Heuristics can be seen as rules of thumb; for situation x, do action y, regardless of the specifics of the situation. Heuristics are "designed to get the response into the right ballpark when solving a problem or making a decision" (Stanovich, 2011, p. 29). It automatically selects the response that comes to mind naturally without consciously going over multiple options.

System 1's heuristic processing has the advantages of being fast, cognitively cheap, meaning it limits the expenditure of cognitive energy. At the same time, it has a high capacity and is (quasi-)modular, which means that "many Type 1 processes can operate at once in parallel" (Stanovich, 2011, p. 29). Therefore, it is highly efficient and not demanding.

As a result of its heuristic nature, system 1 is "systematically biased" (Baron, 2014, p. 4). Heuristics ignore the full scope of information and use the familiarity and known features of the context to reach a decision. Examples of documented biases are confirmation bias (giving undue weight to information which confirms beliefs); anchoring bias (giving undue weight to the first piece of information that is encountered in a decision-context); availability bias (giving undue weight to information that is easily available cognitively); and in-group bias (giving undue weight to information presented by agents from a familiar social group).

6.2.2 Evaluating system 1 thinking

Despite its tendency to engage in biased thinking, system 1 thinking is not inherently undesirable. Some authors emphasize the strengths of system 1 (Gigerenzer, 2007) and system 1 is vital for humans to function efficiently.

As we have described, system 1 is automatic, it is the default way of decision making. It operates on autopilot as the most energy efficient mode of making decisions. Beyond mere cognitive economy, however, system 1 offers significant advantages that are often overlooked. It represents a sophisticated form of pattern recognition and intuitive expertise that can outperform deliberative reasoning in certain contexts. System 1 enables experts to make rapid, accurate judgements in their domains (Kahneman and Klein, 2009). Chess grandmasters, for instance, can recognise board patterns and make strong moves almost instantaneously—a capability that system 2 deliberation alone cannot replicate.

Moreover, system 1 allows people to make good, or good enough decisions based on incomplete information. It excels at navigating environments with inherent uncertainty where system 2 may falter due to computational complexity or missing data. System 1 decision making works particularly well in familiar and favourable environments (Stanovich, 2011). In these contexts, heuristic thinking is not merely efficient but often highly effective and reliable. For example, experienced car drivers do not need to consciously think while driving and can easily manage to have a conversation simultaneously—a testament to the power of automatic processing.

Additionally, system 1 contributes uniquely to social cognition by enabling the rapid processing of facial expressions and social cues—capabilities that deliberative processes cannot match in speed or intuitive accuracy (Lieberman, 2007; Frith & Frith, 2008). This demonstrates that system 1 is not a primitive form of thinking but a different mode of cognition with distinctive strengths.

Despite the advantages that system 1 provides, the nudging literature has predominantly focused on the vulnerabilities of system 1. Saghai's characterization of system 1 as a shallow form of cognition is shared by Stanovich (2011), as well as by most authors in the literature on nudges including Thaler and Sunstein (2008), who describe humans as "mindless, passive decision makers" (Thaler & Sunstein 2008, p. 37). This negative portrayal appears throughout behavioural economics research, where system 1 is consistently depicted as responsible for the errors and biases that underlie the entire behavioural economic framework. The emphasis on these shortcomings serves to justify interventions in decision-making environments, with less attention paid to understanding when and how system 1 thinking might be beneficial or even superior to deliberative processes.

6.2.3 System 2 thinking

Conversely, digital nudges circumvent system 2, which can be characterized as robust cognitive processes. System 2, “allocates attention to the effortful mental activities that demand it” such as computing the answer to 23 x 57 (Kahneman, 2011, p. 21). System 2 processing is experienced consciously and deliberate, it requires attention as well as concentration. As a result, its processes “are often associated with the subjective experience of agency” (Kahneman, 2011, p. 21) and cost more effort than system 1.

As it requires conscious attention, system 2 is serial, meaning it can only process one task at a time (Evans & Stanovich, 2013). This means it is a much slower and more capacity limited form of processes. It is also “computationally expensive” (Stanovich, 2011, p. 79).

While system 1 relies on intuition and past experience, system 2 operates independently of direct experience. It enables individuals to mentally simulate choices, decisions, and contexts before acting. Unlike system 1, which is fast and

automatic, system 2 is more abstract, rule-based, and often dependent on language (Evans & Stanovich, 2013).

Keith Stanovich argues that system 2 must be subdivided into an "algorithmic" system and a "reflective" system (2011, p. 41). The algorithmic system is associated with raw intelligence as typically measured by IQ tests. The reflective system on the other hand is not linked to intelligence but rather to "thinking dispositions" or "cognitive styles" (Stanovich, 2011, p. 41). These dispositions involve someone's tendency to use metacognitive thinking and to engage in reflection. Both are required for making sound decisions.

Proficiency in either the algorithmic or the reflective system is relatively independent of proficiency in the other (Stanovich, 2011). Yet, the algorithmic system is subordinate to the important elements of the reflective system, namely epistemic thinking dispositions such as how beliefs must be acquired and updated in light of new information and goal hierarchy. Consequently, regardless of their innate intelligence level, individuals can develop proficiency in optimally utilising their cognitive capabilities. For example, Robson (2019) offers examples of highly intelligent people who use their cognitive ability to craft nifty arguments for mistaken beliefs, such as the validity of astrology. Such people are intelligent yet have faulty thinking dispositions. They have a strongly developed algorithmic system and a poorly developed reflective system.

As this example illustrates, the distinction between the algorithmic and reflective systems is useful because it allows to "explain unique variance in rational thinking" (p. 40). Rationality requires having good beliefs and acting on them appropriately considering one's preferences. To do so, one must not only have algorithmic system thinking but also be willing to engage in open-minded thinking, consider future consequences, reflect on one's beliefs and goals, and eschew superstition and dogmatism and embrace nuances (Stanovich, 2011). Such thinking dispositions are necessary for adequate decision making.

6.2.4 Evaluating system 2 thinking

Like system 1, system 2 is neither inherently good nor bad. Overuse of the effortful system 2 would be ineffective and lead to cognitive overload. Yet, the literature on nudging has emphasized the benefits of using system 2, citing examples in which deliberation leads to better results, most notably in financial domains where explicit calculation usual yields the best results.

System 2 thinking provides several distinct advantages that complement the limitations of system 1. First, it allows for abstract reasoning and hypothetical thinking, capabilities that are essential for long term planning, complex problem solving, and navigating novel situations (Evans & Stanovich, 2013). This enables agents to consider alternatives that are not immediately present in their environment and to evaluate consequences that extend beyond the immediate future.

System 2 also excels in contexts requiring precision and accuracy. While system 1 provides rough approximations, system 2 can perform exact calculations and systematic analyses when needed. This precision becomes especially valuable in domains such as financial planning, technical design, and statistical reasoning where errors can have significant consequences (Kahneman, 2011).

Furthermore, the reflective component of system 2 enables metacognition, thinking about one's own thinking, which supports learning from mistakes, self-correction, and epistemic improvement over time. This reflective capacity allows individuals to recognize when their intuitive judgments might be misleading and to deliberately override these judgments when appropriate (Stanovich, 2011).

However, system 2 thinking also has notable limitations. Its deliberate, serial nature creates a bottleneck in information processing that can be problematic in time sensitive situations or when managing multiple tasks simultaneously. The high cognitive demands of system 2 processing lead to

mental fatigue and can deplete cognitive resources needed for other tasks (Baumeister et al., 2008).

Moreover, system 2 is vulnerable to its own distinct biases. For instance, overthinking can sometimes lead to worse decisions than intuitive judgments, particularly in domains where expertise has been developed (Dijksterhuis & Nordgren, 2006). System 2 can also enable rationalization, the construction of seemingly logical arguments to justify conclusions reached through system 1 processes or to defend existing beliefs (Mercier & Sperber, 2011).

In educational contexts, overemphasis on system 2 approaches may inadvertently prevent the development of intuitive expertise that comes through repeated practice and the internalization of patterns (Kahneman & Klein, 2009). There is evidence that in some domains, particularly those involving pattern recognition and skilled performance, experts benefit from allowing well trained intuitions to guide their actions rather than engaging in explicit step by step reasoning (Kahneman & Klein, 2009). We elaborate on the importance of system 1 for acquiring new skills and habits in chapter 7.

6.2.5 Triggering vs Subduing

Since system 1 and system 2 each have their own advantages, decision-making requires selecting the most appropriate mode of thinking. This selection is a trade-off "between power and expense" (Stanovich, 2011, p. 41). While system 1 has the advantages of being frugal in terms of computational power, being fast, operating without conscious attention, and performing multiple tasks at the same time. However, it can err due to its heuristic nature and is less powerful. Its task it to "get the response into the right ballpark" (Stanovich, 2011, p. 29). System 2 on the other hand, while being slower and more effortful, can solve problems with greater accuracy when more detail in analysis is required.

When faced with a decision, system 1 responds by default to save concentration and effort. As a result, most decisions agents make are performed

by system 1. One of system 2's crucial tasks it to override system 1 when it appears to go wrong or when the situation is of great importance. Metaphorically, people are on auto-control, reacting intuitively and automatically, until they encounter a novel situation for which there are no learned associations, or a situation that involves great complexity or importance. In that case, system 2 intervenes and overrules the automaticity of system 1.

So, in the ideal scenario system 1 makes most of the decisions. Only when extra attention is required, or when there is a potential conflict of preferences or unclarity in a person's preference hierarchy, system 2 should interfere. In that case, deliberation can be used to assess the full set of benefits and drawbacks of a decision. In novel or unclear situations, agents can use reflection to resolve conflicting goals or update their preference structure.

As a result of this default-intervention process, we argue that the definition of digital nudges as triggering shallow cognitive processes is somewhat misleading. Shallow cognitive processes do not need to be triggered, they are the default automatic form of making decisions. What makes digital environments hostile and digital nudges effective is that they prevent system 2 from intervening. Digital nudges often work by removing all forms of friction (see chapters 3 and 5) that would trigger system 2 to overrule system 1 and cause the agent to adopt a more conscious and deliberate decision-making mode. Hence, digital nudges do not trigger shallow cognitive processes, rather they leverage and utilize the automaticity of shallow cognitive processes, which are the default way of operating.

We must think of digital nudges as operating in a subtle, unseen manner, adeptly manoeuvring the minds of users. Instead of triggering system 1, nudges instead subdue the triggering of system 2's more conscious and deliberate reasoning. When successful, agents are not fully aware of digital nudges operating in stealth mode. This approach is key to their effectiveness. Nudges gently steer choices without rousing the deliberate, analytical scrutiny of system

2, guiding behaviour subtly and unobtrusively. It is a strategy not of confrontation, but of alignment with the default mode of system 1, altering behaviour through delicate suggestion rather than overt persuasion.

6.2.6 Overuse of system 1

Since digital nudges operate by removing the friction that triggers system 2 thinking, they leave agents reliant on the default status, leading to an overuse of system 1.

Overuse of system 1 is particularly problematic in environments that are hostile to heuristics. Recall that heuristics are mental shortcuts which ignore part of the available information. They depend on contextual cues that are adequate for making sufficiently good decisions. Hostile environments contain either few or misleading cues (Kahneman & Klein, 2009). Under these conditions, reliance on heuristics results in misinterpretations and less effective decision-making.

Digital environments are nearly always hostile in this sense because digital nudges control it perfectly. As we saw in chapter 3, digital environments expose agents to a plethora of digital nudges. Many of those digital nudges aim to increase the time agents spend in digital environments by removing all forms of friction, exacerbating the overuse of system 1. Cues are crafted in the interest of the designer, making heuristics unreliable for the agent (Stanovich, 2011). Reliance on system 1 in hostile environments reduces the ability of agents to make good decisions based on the available information, effectively inhibiting their cognitive abilities.

The remainder of this chapter presents three arguments demonstrating how digital nudges impair cognitive abilities by fostering excessive reliance on system 1 thinking, ultimately reducing user welfare.

6.3 System 1 is aligned with lower-order preferences

The first argument is that overuse of system 1 causes agents to give disproportionate weight to lower-order preferences at the expense of their higher-order preferences. Stanovich (2004, p. 36) explains system 1 and system 2 have different goal structures. Since the cognitive mechanisms underlying system 1 are much older in the evolutionary sense, system 1 is aligned more to basic biological needs such as food, shelter, energy conservation and sex. Stanovich refers to these goals as "the gene's interests" (2004, p. 14).

System 1 is direct and immediate. It is geared towards lower preferences which are also immediate, and which provide short-term pleasure. These lower preferences are mostly fixed, as they are determined by biological necessities. They do not differ between humans, and they cannot be changed by will or experience.

Whereas the default implementation of system 1 with its emphasis on basic needs provides an evolutionary advantage to biological organisms, it poses a handicap to agents in pursuit of preferences higher up in the hierarchy.

Higher-order preferences are much more flexible – agents can strive for virtually anything they come to value - they do not follow biological needs as closely. Higher-order preferences are facilitated by system 2. The analytical capacities of system 2 are linked to the prefrontal cortex, the brain's most recently evolved region. Higher-order preferences are a relatively novel phenomenon in biological history, distinguishing humans from other animals (Stanovich, 2004). This is in line with Frankfurt's (1971) argument that what it means to be a person is to have second-order preferences volitions.

System 2 serves long-term preferences which "are utility maximizing for the organism and constantly updated because of changes in environment" (Stanovich, 2004, p. 36). As type 2 cognitive processes are directed more at higher-order preferences, agents are usually better off in terms of utility if they

override the default mechanizations of system 1 when the two systems conflict (Stanovich, 2011, p. 31).

Therefore, Stanovich urges agents to find ways to "transcend the limited interests" of system 1 and "define [their] own autonomous goals" (2004, p. xii). He calls this endeavour a "program of cognitive reform" (2004, p. xii). This argument can be interpreted as stating that agents must ensure they develop the proper cognitive abilities to develop their preferences hierarchies and when they behave in a way that violates their preference structure.

This requires recognizing when a conflict between higher and lower preferences emerges. As system 1 is the default mode of operation, it "will implement [its] short-leashed goals unless overridden" (Stanovich, 2011, p. 44). It is system 2's responsibility to intervene when "inappropriately overgeneralized responses [are] generated by the heuristic system" (Stanovich, 2004, p. 36). When system 1 "suggests" acting on preferences that, as we discussed earlier in the chapter, are lower preferences, system 2 can intervene so that agents can reason about the choice and act on a higher-order preference that is more aligned with their overall welfare.

Therefore, an important cognitive ability is to trigger system 2 when appropriate and use analytical and reasoning to overrule lower-level suggestions from system 1 that conflict with one's higher-order preferences. This ability can be referred to as goal-conflict recognition (Myrseth & Fishbach, 2009). Goal-conflict recognition is one of the two elements of self-control according to Myrseth and Fischbach. When faced with possible temptation, agents must first recognize they are faced with and consider an action that conflicts with their higher-order preferences. Only when agents are aware of the conflict, they can resist the urge to act on the temptation. This second step is referred to as goal-conflict "resolution" (p. 248), which is the topic of the next chapter on willpower. We will here focus solely on the first step.

Saghai argues that the recognition of conflict between preferences "always involves processes that monitor the output of various lower-level cognitive processes to detect a conflict or a mismatch between the agent's explicit or implicit goal and her behavioural response." (Saghai, 2013b, p. 500).

This conflict between system 1's emphasis on lower preferences and system 2's capacity to reflect and pursue higher-order preferences directly impacts the three components of proficient preference shaping established earlier. The overemphasis on system 1 processes created by digital nudges undermines: (1) experiments in living by stimulating choosing immediate gratification rather than diverse experiences; and (2) reflective capacities by bypassing the meta-preference evaluation necessary for proficient preference shaping. The cognitive ability to recognize goal conflicts through system 2 engagement is therefore fundamental to the proficient preference shaping process. Without this recognition, agents cannot engage in the deliberate self-development that Mill identified as essential to welfare. Digital nudges that exploit system 1 processes thus threaten not merely isolated choices but the entire process of preference shaping that is the foundation of welfare.

6.4 Acting for the wrong reasons – cognitive biases

Overuse of system 1 can cause agents to act for the wrong reasons because system 1 thinking relies on heuristic processing, which is biased by nature. Recall that heuristics allow agents to act cognitively efficiently by ignoring a large part of the available information and instead rely on a limited number of informational cues that carry extra weight. Such reasoning makes reasoning biased in the sense that this type of reasoning over relies on predictable patterns. As digital nudges increase the use of system 1, they cause an increase of biased thinking, which hurts welfare. We now turn to the most important biases that are aggravated by digital nudging and how they hurt welfare.

6.4.1 Myopia and hyperbolic discounting

An important and frequently emphasized bias in the behavioural economic literature is myopia, the tendency to focus on immediate or short-term benefits at the expense of incurring larger long-term costs or foregoing larger future benefits. It is common in the economic literature to model intertemporal choices with "time preferences", which represent the extent to which agents exhibit a preference for receiving benefits sooner rather than later, at a future cost.

In the literature on nudging, overvaluing the present is usually interpreted as a mistake. One of the most prominent examples Thaler and Sunstein (2008) provide is the lack of money that American households save. They argue that the high consumption and low personal savings rates, as well as the low pension contributions, create financial trouble for people. According to them, the lack of financial frugality means "folks could clearly use a nudge" (p. 107), which shows they interpret the lack of saving as a mistake.

More formally, Thaler points out that agents exhibit time-inconsistent preferences when they are more willing to postpone a benefit from one week to two weeks compared to postponing a benefit from now to one week (Thaler, 2015). Such choice behaviour, which has been found in many studies (see for example Hoch & Loewenstein, 1991; Grenadier & Wang, 2007) means agents not only value present benefits more than future benefits but also discount a week in the near future at a different rate from a week in the very near future. Discounting occurs when agents adjust how much they value some outcome based on how far into the future that outcome is. But when agents discount time in the near future at a very different rate from the same amount of time in the far future, this is a form of hyperbolic discounting which leads to "time-inconsistent behaviour" (Thaler, 2016, p. 1592).

According to Thaler and Sunstein, the lack of saving is at least partially caused by policies that are based on neoclassical economic theory, which

assumes agents calculate how much savings they need later in life and diligently act on this calculation. In their view, these policies do not work because first, people lack the calculative capabilities to make a sound plan and second, lack the willpower to stick to the plan (Thaler & Sunstein, 2008, p. 104). We discuss willpower in the next chapter. Here, we focus on the lack of capabilities.

An appropriate valuation of the future requires a valuation of the immediate benefits, the future benefits, and a sensible discount rate. System 1 offers a rapid ballpark comparison of the immediate and easy to obtain benefits. More concrete, salient, visible and direct benefits carry more weight in the comparison. Overuse of system 1 would hence give a disproportionately large weight to features with these characteristics, overvaluing the present.

Our point is not that agents must make some perfect calculation to optimize their rate of saving. Rather, we want to emphasize that overreliance on System 1 diminishes opportunities for reflection and conscious decisions which limits effective decision-making. Particularly, it limits an agent's ability to consider benefits that are more abstract or far away in time.

As a result, overusing system 1 leads to a loss of welfare because it overvalues the present and undervalues the future. Overuse of system 1 inhibits the appropriate evaluation of the importance of the long-term benefits when making a choice. Agents will improve the quality of their choices if they have the capacity to involve system 2 in choices that affect their long-term future. Digital nudges that succeed in keeping the agent in system 1 type thinking will create this result.

While most examples of nudges mentioned in the literature on the ethics of nudging involve either money or health (which have interesting applications in digital environments), we want to point out another domain that is particularly interesting from a digital perspective, namely how agents spend their time. In chapter 3, we have compared algorithmic timelines with slot machines.

Mindlessly scrolling an algorithmically sorted timeline can be interpreted as an endless offer of immediate benefits, in the form of small dopamine boosts. While both the marginal benefit and marginal cost of viewing an additional post are small, the cumulative cost can be large as they consume a great amount of time and attention. We have noted before such timelines are completely optimized to capture time and attention. We also highlighted that in social media apps, all friction is eliminated to ensure users stay in type 1 thinking, preventing them from reflecting on their behaviour and clicking away.

The result is that people self-report they spend too much time on their phones (Schuster, Cotton & Meshi, 2022). With digital activities such as streaming, gaming and scrolling on social media, people report an urge to continue and sometimes an inability to stop (Holmgren & Coyne, 2017). We have touched on issues of addiction in an earlier chapter.

Moreover, it is important to note what time on social media and digital environments replaces. According to Jonathan Haidt, childhoods have made the "transition from a play-based childhood involving a lot of risky unsupervised play, which is essential for overcoming fear and fragility, to a phone-based childhood which blocks normal human development by taking time away from sleep, play, and in-person socializing" (Haidt, 2023).

By keeping people in type 1 cognition, digital nudges make agents overly sensitive to desire one more click and fall for instant gratification, neglecting future benefits of spending their time otherwise.

While we review resisting the urge to instant gratification in the next chapter on willpower, it is important to highlight here that system 1 thinking has a bias towards quick rewards, without accounting for future costs and benefits. Accounting for the future is a cognitively more high-level task which requires system 2 to intervene. Agents who overuse system 1 hence are overly sensitive to short term benefit hoarding, hurting their own long-term welfare.

6.4.2 Cognitive fluency

The second bias, or rather collection of various biases and heuristics, is the tendency to give disproportionate weight in one's thinking to informational cues that are cognitively fluent. Cognitive fluency is the ease with which information is processed. Those informational elements that are processed easily are considered to be truer and influence the evaluation of information more strongly than informational that is cognitively hard to process.

When information is processed fluently, it is more likely to be interpreted as true, valuable, and relevant (Oppenheimer, 2008). This shapes preferences as agents prefer things that are easy to understand and even have cognitively fluent names (Constable *et al.*, 2013). For example, stocks with fluent or easily pronounced names outperform stocks with complex names (Alter & Oppenheimer, 2006).

Digital nudges that succeed in keeping agents in system 1 thinking make agents susceptible to overlying on information that is cognitively fluent because system 1 is, as noted before, a cognitive miser. Agents who process information using lower cognitive processes seek to save as much cognitive effort as possible. When they need to process information, they try to find the easiest available answer. If that answer feels right, they accept it. This leads to several biases.

Availability bias is the tendency to rely on recent examples and experiences in one's thinking. When evaluating a choice or opinion, more recent experiences tend to come to mind more quickly, be more cognitively fluent, and have a disproportionate influence. For example, respondents are likely to greatly overestimate the probability of aircraft crashes after stories of incidents with aircrafts in the media because when they are asked about probabilities, they do not actually assess statistical evidence but the ease with which they can recall examples (Tversky & Kahneman, 1973).

In addition to availability bias, there is evidence for the existence of familiarity bias, or mere-exposure effect, the tendency to prefer familiar options

over unfamiliar ones without regard for their qualities. The more familiar something is, the more cognitively fluently it is processed, and the more likely it is to be perceived positively. This way, familiarity shapes affect and preferences (Constable et al., 2013). The application of this idea is the basis of many marketing and political campaigns where repeated exposure to a message creates familiarity and increases its acceptance.

Finally, there is a bias commonly referred to as "what you see is all there is (WYSIATI)" (Kahneman, 2011, p. 73). This refers to the tendency to base decisions and judgements on the information that is available without considering which relevant information might be missing. This bias "facilitates the achievement of coherence and of the cognitive ease" by skipping the effort of assessing which information might be missing (Kahneman, 2011, p. 75). This bias ensures agents are too likely to base decisions on information that is presented to them, and not likely enough to search for new information they might need to make a good decision.

The combination of these biases means that when overusing system 1, agents give disproportionate weight to information that is cognitively fluent, including information that they learned recently, is familiar, and is simply present.

These biases can be supplemented by a range of other factors that facilitate cognitive fluency such as ease of navigation and display (Arning & Ziefle, 2009), readability of colours and fonts (Song & Schwartz, 2008), and simply worded information (Oppenheimer, 2006).

If agents overuse system 1, the combination of these biases harms their welfare because it causes them to give disproportionate weight to what is recent, available, easy, familiar and present. Conversely, agents do not engage enough with the new, complex and unfamiliar. An example in the digital world that illustrates the harmful welfare effects of digital nudges keeping agents in system 1 thinking is the autoplay feature on social media platforms like YouTube or

Netflix. The autoplay feature facilitates cognitive fluency by allowing the user to continue watching shows, even after an episode is finished and even when a series is finished. As a result, all friction is removed from the process of deciding if and what to watch, keeping users in system 1 type thinking.

The algorithm that determines which video will be (auto)played next feeds into the availability, familiarity biases. By suggesting content that is close to what the user already knows and understands, the algorithms limit conscious awareness and prevents system 2 processes from interfering.

This has potentially harmful welfare effects, because it limits the user to discover new and potentially more rewarding content. In this way, the autoplay feature limits the diversity to which the user is exposed. In addition, the autoplay feature keeps intellectually stimulating content from the user because more superficial content is more likely to keep users in system 1 and therefore optimizes screen time.

The autoplay feature also exploits the wysiati bias by limiting the choices that are visible to the user, preventing the user to look for more challenging or rewarding content. The wysiati bias reinforces the other two biases as the autoplay feature not only feeds content that is available, easy, and familiar but also limits the visible choices.

Moreover, the recommender algorithm in the autoplay feature has the ability to learn and will promote videos that increase watching time while ignoring videos that decrease watching time. This means that if someone for example watches a YouTube video on learning a hand skill, and the video stimulates the user to start practicing that skill, the algorithm will punish that video as it causes users to leave the platform. It thus actively discourages users to engage with content that stimulates action.

Our point is not that users should avoid platforms such as YouTube and Netflix, the point is that after finishing a video, it would be best to have a small moment of reflection to evaluate if the previous video delivered what was hoped

for. Then, the user could quickly reflect on whether they want to watch another, and if so, what type of video. This small moment of reflection should at least consider less cognitively fluent or easy pieces of information.

The autoplay feature shows how the interplay of biases that prefer cognitive fluent information causes users to be overly likely to engage with superficial and familiar information. It also means users miss important opportunities to engage with novel, diverse or complex content, hurting the user' ability to engage in experiments in living and ultimately their welfare.

6.4.3 Risk Aversion

An important and well-researched cognitive bias is risk aversion - the tendency to avoid risk and prefer certainty (Andreoni & Sprenger, 2012). Risk aversion is not to be confused with loss aversion, which is the tendency to respond "more strongly to losses than to gains (Kahneman, 2011, p. 91). An agent's choice between a 50% chance of winning $400 and a 100% of winning $200 shows an agent's preference for risk while neither involve losses. While risk aversion is often measured through monetary experiments for ease of quantification, its principles are broadly applicable to any situation involving risk (Kahneman & Tversky, 1984). For example, Kovářík and van der Leij (2014) found that risk averse individuals have closer social networks consisting of people they are familiar with.

While many people are risk averse when using system 2 as well, risk aversion is more pronounced in system 1 thinking (Kahneman, 2011). Using system 1 reasoning leads to more emphasis on safety and certainty. When relying on system 1 cognitive processes, agents tend to prefer certainty and are less drawn to risks.

Yet, as we have argued in chapter 4 based on Mills thinking, experiments in living are a key ingredient for the proficient shaping of one's preferences thereby for welfare. Experiments in living are attempts to find value in new and

unknown territory. The rewards of performing such experiments are all but certain and hence involve risk for the experimenter.

So, when agents overuse system 1, they are reluctant to engage with novel activities that are unsure to be rewarding and will hence be unwilling to engage in the experiments in living that are necessary for welfare on both an individual and societal level, as explained in chapter 4.

The cognitive bias of risk aversion works in tandem with the overload of negative information discussed in the previous chapter. On the one hand there is more information available about risk. On the other hand, people are affected by biases that affect their interpretation of available information and amplify their perception of risk. Together, these two mechanisms are at the basis of what Kasperson et al. call "the social amplification of risk" (1988, p. 177), which concept they use to describe the social practices surrounding risk, including adverse consequences such as overregulation, over insurance, low levels of technology adoption and "[e]nduring mental perceptions, images, and attitudes (e.g., antitechnology attitudes, alienation from the physical environment [and] social apathy)" (p. 182).

While Kasperson et al. (1988) already pointed to the media as an important determinant of risk perception in society, the effects of social media on risk perception have been researched more recently. Specifically, the effects of social media on risk perception during infectious disease outbreaks (Oh & Lee, 2021). A set of studies covering the 2015 MERS outbreak in South-Korea (Oh & Lee, 2021; Song et al., 2017; Ofoghi et al., 2016) finds evidence that on social media, responses to the outbreak were emotional, with the emotions of fear and anger being most prevalent. Oh & Lee (2021) find that social media use was positively associated with experience of fear and anger, as well as with risk perception.

More generally, there is evidence that social media use causes symptoms of depression and anxiety, especially in teens and adolescents. Keles, McGrae &

Grealish (2020) conducted a systemic review of 13 studies on the effects of social media use on the mental health of adolescents. They investigated four domains: time spent, activity (active contribution versus passive scrolling), investment (effort put into it, e.g. optimizing social profiles) and addiction (dependency), and found that all four "correlated with depression, anxiety and psychological distress" (p. 79). An important caveat is that the link between social media use and mental health problems is complex and possibly mediated by various factors such as sleep deprivation, lack of physical exercise, personal traits, sociocultural factors and cyberbullying (Keles, McGrae & Grealish, 2020).

The adverse mental effects of social media use cannot be completely attributed to risk aversion. It is also related linked to how people and particularly children spend their time (as discussed), social pressure (next section) and other factors (which we cannot fully discuss here). However, the combination of extensive exposure to negative and emotional content, as discussed in the previous chapter, with an overreliance on system 1 thinking, makes individuals more susceptible to risk aversion and interpreting risk information disproportionately. This fosters fear and discourages experimentation and diverse experiences, ultimately harming overall welfare.

6.4.4 Conformity

Conformity bias is the tendency to base decisions or beliefs to conform with others (Cialdini, 1984). This phenomenon is also referred to as follow the herd tendency, social proof, or social influence. In the previous chapter, we discussed how this bias is linked to the over-prevalence of social information, promoted by algorithmic information filters. Conformity bias and the over-prevalence of social information reinforce each other, amplifying the effects of norms and the behaviour of others on individual behaviour. Here, we want to go into how conformity bias is aggravated by overuse of system 1, and how this overuse harms welfare.

People are naturally likely to seek to fit in with others around them. When overusing system 1, this tendency is more prevalent as system 1 is more automatic and devoid of conscious reflection. Agents reflect less on the behaviour of others and may copy it more mindlessly. So, overuse of system 1 aggravates the strength of conformity bias, retendering the behaviour of others a stronger factor of influence.

The increased influence of the behaviour and opinions of others can stifle individuality, reduce critical thinking, and limit diverse perspectives, ultimately impacting personal and societal welfare. In chapter 4, we outlined Mill's concept of welfare and how the notion of individuality is "one of the leading essentials" for it (Mill, 2001/1859, p. 53).

To attain individuality, agents must not only be free from coercion but also actively form their own preferences without too much regard to custom and norms. As Mill (2001/1859, p. 53) formulates it: "[i]t is desirable, in short, that in things which do not primarily concern others, individuality should assert itself. Where not the person's own character, but the traditions or customs of other people are the rule of conduct, there is wanting one of the principal ingredients of human happiness, and quite the chief ingredient of individual and social progress." So, agents should not be too quick to conform to the dominant norms and expectations of others. To attain welfare, agents must discover for themselves through experimentation and experience what they value. Systematic overuse of system 1 creates a more mindless mode of living and amplifies conformity bias, hindering the individuality that is necessary for welfare.

Social norms and the opinions of others are most visible on social media. Social media are characterized by large amounts of information; a large amount of social information as defined in the previous chapter, interactions with a great amount of people, notably more than would be possible in real life; explicit validation mechanisms such as likes or hearts. In this way, social media users are exposed to rapid feedback from a large group of others, including many they

do not know (well) (Blomfield Neira & Barber, 2014). As a result, social media are a hotbed for "social comparison and peer feedback" (Blomfield Neira & Barber, 2014, p. 56).

The overuse of system 1 which harnesses conformity bias in combination with the large amount of social information is especially damaging for adolescents, whose identities are still fluid and under development. Adolescents use social comparisons and feedback from peers for self-evaluation (Harter, 1999).

Making social comparisons on social media can be especially confronting because, other than in "real life", social media users have great control over what they share about themselves yet little control over what they see. As a result, social media are full of success stories, causing the perception that others "are happier and having better lives" (Chou & Edge, 2012, p. 117).

Another reason social comparisons are facilitated by social media the visual nature of social media, which facilitates salience. Especially when overusing system 1, a picture has a much greater emotional effect than a written story. Being confronted not only with the knowledge that others are doing something "better", but also with visual proof of those activity can enhance social comparisons and feelings of envy.

Together, these characteristics mean social media are "drowning kids in social comparisons they can't win" (Haidt; afterbabel.com, 2023). The overuse of system 1 and the resulting over-reliance on conformity bias harm the individuality that is necessary for welfare. Especially adolescents, who are forming their identity and preferences to a greater extent than adults, suffer from excessive comparison to others.

This is highlighted by a broad range of evidence that links social media use to increased mental problems, anxiety and depression in teens and adolescents (Keles, McGrae & Grealish, 2020). One study reports a natural experiment by comparing mental health symptoms of students before and after Facebook was

introduced at their college - using the phased introduction of Facebook across colleges. They found a negative effect on mental health and depression after the introduction of Facebook. The authors note that "evidence on mechanisms suggests the results are due to Facebook fostering unfavourable social comparisons." (Braghieri, Levy, & Makarin, 2022). Appel et al. (2016) also link Facebook use to depression and "negative affectivity" (p. 1). They also mention the role of social comparisons and point to the relationship between envy and depression.

These examples and studies show how digital nudges can harm welfare by ensuring agents overuse of system 1 and thereby over-rely on their conformity bias. This bias means agents, continuously compare themselves to others. This is further enabled by social media sites and the disproportionate availability of social information as discussed in the previous chapter. Together, these mechanisms hamper the individuality that is necessary for discovering and acting on those things one finds valuable, no matter what others think of it.

6.5 Long-term undermining of system 2 abilities

So far, we have discussed the harms of overusing system 1 at the expense of using system 2. Now, we turn to the specific harms to welfare caused by the long-term deterioration of system 2 abilities. A long-term perspective on cognitive abilities highlights that the adverse effects of digital nudges may persist even when digital nudges are not present.

This argument shares similarities with the argument of infantilization, which we defined in chapter 2 as follows: agents who are subject to many nudges over a longer period will develop eroded autonomous decision-making capabilities and will lack "responsibility in matters regarding [their] own welfare" (Bovens, 2009: 11). Dold & Schubert similarly argue "nudges may impair agents' 'individuation' if they systematically discourage active choice –

understood as deliberate and effortful decision-making and the exercise of self-control." (Dold & Schubert, 2018, p. 234). While Bovens as well as Dold and Schubert's comments mainly refer to libertarian paternalistic nudges, the worry can be extended to private and digital nudges.

In chapter 4, we defined evaluative and reflective capacities as one of the three elements of the skills needed for proficiently shaping preferences and experiments in living. Evaluative and reflective capacities are necessary to reflect on one's preference structure and to weigh diverse, potentially conflicting, preferences against each other.

As system 1 operates on default and is drawn by lower preferences, agents must recognize moments in which they are tempted to act against their higher-order preferences as well as situations in which their preferences are not yet fully formed. In addition, agents must be able to reflect on which preferences matter most to them, and which direction they would ideally want their preferences to develop "character or system of preferences to develop over time" (McPherson, 1982, p. 260). Doing so requires the evaluative and reflective abilities of system 2. It requires what we called the "reflective system" (Stanovich, 2011, p. 41). A well-developed reflective system allows one to engage in open-minded thinking, consider future consequences, reflect on one's beliefs and goals, and eschew superstition and dogmatism and embrace nuances (Stanovich, 2011).

These cognitive abilities of the reflective system are best thought of as skills. As Julia Annas writes in her book *Intelligent Virtue*, "the skilled person can 'give an account' of what he does, which involves being able to explain why he is doing what he is doing" (Annas, 2011, p. 20). For her, virtues such as wisdom are akin to skills. It is a skill to have reasons for one's actions and be aware of them.

Skills must be practiced to fully develop them. There is evidence that actively open-minded thinking can be trained (Lambie, 2014). In addition, there is evidence that such training reduces the effect of cognitive biases such as

anchoring (Adame, 2016) and overconfidence (Hirt, Kardes & Markman, 2004). So, developing reflective and evaluative capacities takes training and practice.

Underuse of these skills results in lack of practice and a deterioration of skill. The flip side of the overuse of system 1 is the underuse of system 2's reflective capacities. Digital nudges ensure agents underuse their reflective capacities and therefore fail to train them. Hence, digital nudges hurt reflective thinking which is one of the elements needed for proficiently shaping preferences.

To provide an example, consider someone scrolling mindlessly on social media for hours, even though they wanted to work on an important project. When waking up from this automatic and mindless state, mindless scrollers will realize they did not act in accordance with their own goals, and therefore acted against their higher-order preferences, and will as a result experience cognitive dissonance. In this moment, mindless scrollers can ease the cognitive dissonance, which is an uneasy feeling in two ways. Either they seek to justify their behaviour by crafting some sort of justification for his behaviour such as: there will be time tomorrow to work on the project, or they try to open-mindedly reflect on why they acted against their preferences. Such open-minded reflection on one's own preference is difficult, yet it is a skill that can be learned.

6.6 Recommendations

To mitigate the harmful effects of digital nudges on cognitive abilities and thereby welfare, we propose the following recommendations that are targeted at high schools and educational policy makers and are structured similarly to the previous chapter. By focusing on strengthening critical thinking skills and the cognitive abilities needed for proficient preference shaping, we aim to better equip students to navigate the challenges posed by digital nudges.

6.6.1 Integrate theoretical knowledge about digital nudges and cognitive abilities into curricula

To limit the effects of digital nudges on behaviour and cognitive abilities, students must be made aware of how digital nudges operate. Minimally, students should be made aware of the two thinking modes we referred to as system 1 and system 2. For teaching purposes, the more memorable terms “hot” thinking versus “cold” thinking could be used (Pekel; Search Institute, 2016). When students understand these two modes and recognize that both have their own strengths and weaknesses, they can label which mode they are operating in and reason critically about how it affects their decision-making. Particularly, educators could emphasize that being in the hot thinking state (system 1), makes one more likely to act on lower preferences and more likely to do things one regrets later. In addition, teachers could emphasize that hot thinking reacts more strongly to cues in the environment. Combined with the awareness that digital designers structure their digital environments, as discussed in the previous chapter, this would be a good first step to understanding the potential harms of digital nudges.

In addition to awareness of the two thinking modes, students should learn common cognitive biases such as the myopia, cognitive fluency, risk aversion and conformity biases. Teachers could discuss practical examples of these biases that resonate with students and apply cold state thinking to biased conclusions. For example, they could discuss how financial influencers seem credible due to use of status symbols and subject their financial claims to more empirical rigor in classrooms.

6.6.2 Develop programs focused on strengthening cognitive abilities and reflection

In addition to cultivating awareness, education programs should help students recognize that cognitive abilities can be greatly improved, even though popular belief depicts intelligence and cognitive abilities as fixed. If the quality of one's decisions were merely a reflection of their innate talent and inherited qualities, few recommendations concerning the cognitive abilities of agents would be fruitful and we would be limited to improving cognitive abilities indirectly through manipulation of the environment. However, research suggests there is ample room for direct improvement of cognitive abilities as well. Encouraging both educators and students to adopt this mindset can help students believe in their capacity to enhance their cognitive abilities. In addition, belief in progress helps students take control of their own lives, engage in experiments and be open to new experiences.

The conceptual move we made between the algorithmic system and the reflective system creates room for this improvement. This distinction is helpful because the reflective system is not linked to intelligence but to thinking dispositions. In contrast to raw intelligence, thinking dispositions, such as engaging in open-minded thinking, considering future consequences, reflecting on one's beliefs and goals, eschewing superstition and dogmatism, and embracing nuances, can be greatly improved by practice. Thus, while IQ remains relatively stable throughout life, individuals can develop their metacognitive skills to utilize their intelligence more effectively.

The research field of boosting seeks to improve thinking dispositions, so people make sound decisions and use the information available to them optimally. While the nudge approach assumes "somewhat mindless, passive decision makers" (Thaler and Sunstein 2008, p. 37), whose automatic system 1 can be manipulated, the boosts approach "assumes a decision maker whose

competences can be improved by enriching his or her repertoire of skills and decision tools" (Grüne-Yanoff & Hertwig, 2016, p. 152). In this way, boosts aim to improve agent proficiency in making decisions and exercising their autonomous agency (Hertwig & Grüne-Yanoff, 2017).

6.6.3 Demonstrate techniques for introducing friction into digital interactions

Having identified overuse of system 1 as the main culprit for cognitive abilities, students must develop ways to limit the system 1 use and activate system 2. As explained above, friction causes system 2 to become alert and intervene. Where digital nudges remove all forms of friction, students could learn to reintroduce friction. Education should teach various ways in which students can do this. This can be done through psychological mechanisms, such as learning associations with certain actions. For example, a person could train themselves to, upon entering the house, immediately place their smartphone in a secure, hard-to-reach spot, out of sight. If students developed this habit, they would be less likely to reach for their smartphones and therefore less susceptible to digital nudges. This approach could also be incorporated into educational settings, where some schools already require smartphones to be stored in lockers during school hours (Singer, 2024).

Another way to introduce friction is through digital tools, as discussed in an earlier chapter. These tools usually work by either blocking information, such as blocking the YouTube sidebar, or by giving a warning after an extended period of use. Schools could teach students on which digital tools are available and how they impact behaviour.

These efforts help trigger system 2 with its more reflective nature. In this way, students can be thought to prevent overuse of system 1, use system 2 more often, and reflect on their own digital behaviour more often and more deeply.

6.6.4 Practice critical thinking tools in digital contexts

Kenyon & Beaulac (2014) investigated whether providing students with knowledge of the existence of cognitive biases was sufficient for reducing their effects, a belief that was widespread. They concluded that this idea was "highly over-optimistic" at best (Beaulac & Kenyon, 2018 p. 94). Instead, educators should teach students "debiasing" strategies that help students to combat the effects of nudges. Debiasing usually takes the form of teaching students general critical thinking to submit judgements to scrutiny. There has been substantial research on how students can be taught to reason better about statistical information. For example, Gerd Gigerenzer set up a teaching program to teach 8-year-olds probabilistic reasoning (Bond, 2009). The idea behind this program is not to steer behaviour of students directly, but to help them make better decisions for themselves. Teaching programs like these can significantly reduce the effects of cognitive biases such as myopia and risk aversion, reducing the harmful welfare effects of digital nudges.

6.7 Bibliography

Adame, B. J. (2016). Training in the mitigation of anchoring bias: A test of the consider-the-opposite strategy. *Learning and Motivation, 53,* 36–48. https://doi.org/10.1016/j.lmot.2015.11.002

Alter, A. L., & Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. *Proceedings of the National Academy of Sciences*, *103*(24), 9369–9372. https://doi.org/10.1073/pnas.0601071103

Annas, J. (2011). *Intelligent virtue.* Oxford University Press.

Andreoni, J., & Sprenger, C. (2012). Risk Preferences Are Not Time Preferences. *American Economic Review*, *102*(7), 3357–3376. https://doi.org/10.1257/aer.102.7.3357

Appel, H., Gerlach, A. L., & Crusius, J. (2016). The interplay between Facebook use, social comparison, envy, and depression. *Current Opinion in Psychology*, *9*, 44–49. https://doi.org/10.1016/j.copsyc.2015.10.006

Arning, K., & Ziefle, M. (2009). Effects of age, cognitive, and personal factors on PDA menu navigation performance. *Behaviour & Information Technology*, *28*(3), 251–268. https://doi.org/10.1080/01449290701679395

Baron, J. (2014). Heuristics and biases. The Oxford Handbook of Behavioral Economics and the Law, 3–27.

Baumeister, R. F., Sparks, E. A., Stillman, T. F., & Vohs, K. D. (2008). Free will in consumer behaviour: Self-control, ego depletion, and choice. *Journal of Consumer Psychology*, *18*(1), 4–13.

Beaulac, G., & Kenyon, T. (2014). Critical thinking education and debiasing (AILACT Essay Prize Winner 2013). *Informal Logic*, *34*(4), 341–363.

Beaulac, G., & Kenyon, T. (2018). The Scope of Debiasing in the Classroom. *Topoi*, *37*(1), 93–102. https://doi.org/10.1007/s11245-016-9398-8

Blomfield Neira, C. J., & Barber, B. L. (2014). Social networking site use: Linked to adolescents' social self-concept, self-esteem, and depressed mood. *Australian Journal of Psychology*, *66*(1), 56–64. https://doi.org/10.1111/ajpy.12034

Bond, M. (2009). Decision-making: Risk school. *Nature*, *461*(7268), 1189–1192. https://doi.org/10.1038/4611189a

Bovens, L. (2009). The ethics of nudge. In *Preference change: Approaches from philosophy, economics and psychology* (pp. 207-219). Dordrecht: Springer Netherlands.

Braghieri, L., Levy, R., & Makarin, A. (2022). Social Media and Mental Health. *American Economic Review*, *112*(11), 3660–3693. https://doi.org/10.1257/aer.20211218

Cialdini, R. B. (2009). *Influence: Science and practice* (5th ed.). Pearson Education.

Chou, H.-T. G., & Edge, N. (2012). "They Are Happier and Having Better Lives than I Am": The Impact of Using Facebook on Perceptions of Others' Lives. *Cyberpsychology, Behavior, and Social Networking*, *15*(2), 117–121. https://doi.org/10.1089/cyber.2011.0324

Constable, M. D., Bayliss, A. P., Tipper, S. P., & Kritikos, A. (2013). Self-generated cognitive fluency as an alternative route to preference formation. *Consciousness and Cognition*, *22*(1), 47–52. https://doi.org/10.1016/j.concog.2012.11.006

Dijksterhuis, A., & Nordgren, L. F. (2006). A theory of unconscious thought. *Perspectives on Psychological science*, 1(2), 95-109.

Dold, M. F., & Schubert, C. (2018). Toward A Behavioral Foundation of Normative Economics. *Review of Behavioral Economics*, *5*(3–4), 221–241. https://doi.org/10.1561/105.00000097

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. The Journal of Philosophy, 68(1), 5–20.

Fuchs, A. E. (2001). Autonomy, slavery, and Mill's critique of paternalism. *Ethical Theory and Moral Practice*, *4*, 231–251.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. Viking. http://www.myilibrary.com?id=705549

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143.

Grenadier, S. R., & Wang, N. (2007). Investment under uncertainty and time-inconsistent preferences. *Journal of Financial Economics*, *84*(1), 2–39.

Grüne-Yanoff, T., & Hertwig, R. (2016). Nudge Versus Boost: How Coherent are Policy and Theory? *Minds and Machines*, *26*(1–2), 149–183. https://doi.org/10.1007/s11023-015-9367-9

Haidt, J. (2023). Social media is a major cause of the mental illness epidemic in teen girls: Here's the evidence. *After Babel*. https://www.afterbabel.com/p/social-media-mental-illness-epidemic

Harter, S. (1999). The construction of the self: A developmental perspective. Guilford Press.

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science*, *12*(6), 973–986. https://doi.org/10.1177/1745691617702496

Hirt, E. R., Kardes, F. R., & Markman, K. D. (2004). Activating a mental simulation mind-set through generation of alternatives: Implications for debiasing in related and unrelated domains. *Journal of Experimental Social Psychology, 40*(3), 374–383. https://doi.org/10.1016/j.jesp.2003.07.009

Hoch, S. J., & Loewenstein, G. F. (1991). Time-inconsistent Preferences and Consumer Self-Control. *Journal of Consumer Research*, *17*(4), 492–507. https://doi.org/10.1086/208573

Holmgren, H. G., & Coyne, S. M. (2017). Can't stop scrolling!: Pathological use of social networking sites in emerging adulthood. *Addiction Research & Theory*, *25*(5), 375–382. https://doi.org/10.1080/16066359.2017.1294164

Song, H., & Schwarz, N. (2008). Fluency and the Detection of Misleading Questions: Low Processing Fluency Attenuates the Moses Illusion. *Social Cognition*, *26*(6), 791–799. https://doi.org/10.1521/soco.2008.26.6.791

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341.

Kasperson, R. E., Renn, O., Slovic, P., Brown, H. S., Emel, J., Goble, R., Kasperson, J. X., & Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk Analysis*, *8*(2), 177–187.

Keles, B., McCrae, N., & Grealish, A. (2020). A systematic review: The influence of social media on depression, anxiety and psychological distress in adolescents. *International Journal of Adolescence and Youth*, *25*(1), 79–93. https://doi.org/10.1080/02673843.2019.1590851

Kiely, K. M. (2014). Cognitive Function. In A. C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 974–978). Springer Netherlands.

Kovářík, J., & Leij, M. J. van der. (2014). Risk Aversion and Social Networks. *Review of Network Economics*, *13*(2), 121–155. https://doi.org/10.1515/rne-2014-0029

Lambie, J. (2014). How to be critically open-minded: A psychological and historical analysis. Palgrave Macmillan.

McPherson, M. S. (1982). Mill's Moral Theory and the Problem of Preference Change. *Ethics*, *92*(2), 252–273.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences*, *34*(2), 57-74.

Mill, J. S. (2001). *On liberty* (Original work published 1859). Batoche Books.

Mill, J. S. (2001). *Utilitarianism* (Original work published 1863). Batoche Books.

Myrseth, K. O. R., & Fishbach, A. (2009). Self-Control: A Function of Knowing When and How to Exercise Restraint. *Current Directions in Psychological*

Science, *18*(4), 247–252. https://doi.org/10.1111/j.1467-8721.2009.01645.x

Blomfield Neira, C. J., & Barber, B. L. (2014). Social networking site use: Linked to adolescents' social self-concept, self-esteem, and depressed mood. *Australian Journal of Psychology*, *66*(1), 56–64. https://doi.org/10.1111/ajpy.12034

Oh, S.-H., Lee, S. Y., & Han, C. (2021). The Effects of Social Media Use on Preventive Behaviors during Infectious Disease Outbreaks: The Mediating Role of Self-relevant Emotions and Public Risk Perception. *Health Communication*, *36*(8), 972–981. https://doi.org/10.1080/10410236.2020.1724639

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Applied Cognitive Psychology*, *20*(2), 139–156. https://doi.org/10.1002/acp.1178

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*(6), 237–241. https://doi.org/10.1016/j.tics.2008.02.014

Pekel, K. (2016). *April Renaissance kit*. Search Institute. http://www.jostensrenaissance.com/wp-content/uploads/2016/04/AprilRenKit.pdf

Robson, D. (2019). The intelligence trap: Why smart people make dumb mistakes. Basic Books.

Saghai, Y. (2013a). Salvaging the concept of nudge. *Journal of Medical Ethics*, *39*(8), 487–493. https://doi.org/10.1136/medethics-2012-100727

Saghai, Y. (2013b). The concept of nudge and its moral significance: A reply to Ashcroft, Bovens, Dworkin, Welch and Wertheimer. *Journal of Medical Ethics*, *39*(8), 499–501. https://doi.org/10.1136/medethics-2012-101112

Schuster, A. M., Cotten, S. R., & Meshi, D. (2023). Established Adults, Who Self-Identify as Smartphone and/or Social Media Overusers, Struggle to Balance Smartphone Use for Personal and Work Purposes. *Journal of Adult Development*, *30*(1), 78–89. https://doi.org/10.1007/s10804-022-09426-3

Singer, N. (2024, August 11). Why schools are racing to ban student phones. *The New York Times*. https://www.nytimes.com/2024/08/11/technology/school-phone-bans-indiana-louisiana.html

Stanovich, K. E. (2004). The robot's rebellion: Finding meaning in the age of Darwin. University of Chicago Press.

Stanovich, K. E. (2011). *Rationality and the reflective mind*. Oxford University Press.

Thaler, R. H. (2015). Misbehaving: The making of behavioural economics. Penguin Books.

Thaler, R. H. (2016). Behavioral Economics: Past, Present, and Future. *American Economic Review*, *106*(7), 1577–1600. https://doi.org/10.1257/aer.106.7.1577

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Penguin.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9.

Chapter 7 – How digital nudges undermine willpower

7.1 Introduction

After information and cognitive abilities, we now turn to the third fundamental limitation: willpower. We analyse how digital nudges diminish agent willpower and thereby their proficient shaping of preferences, which ultimately harms welfare. Willpower is related to numerous benefits such as better academic performance, lower chance of falling into crime, and better personal relationships, even when correcting for intelligence, social class, and race (Baumeister & Tierney, 2012). It also predicts success in terms of income, financial security, status and health (Duckworth, 2011).

Other names for willpower, or close synonyms, include self-control (Ainslie, 2006; Duckworth 2011), self-constraint (Cowen, 1991), self-management (Schelling, 1978), "self-regulation, self-command, self-denial, self-discipline, self-mastery, self-restraint, [...] self-government" (Ainslie, 2021, p. 6), "delay of gratification, effortful control, [...] executive control, [...], self-discipline, self-regulation, and ego strength" (Duckworth, 2011, p. 2639). A common way of thinking about willpower is in terms of akrasia, weakness of will, which can be described in Frankfurt's terms as a conflict between a high-order preference to abstain from a hedonic pleasure and a lower-order preference to engage in it. Akratic behaviour can be viewed as the temporary victory of the lower-order preferences.

While the fundamental limitation that agents face is a lack of willpower, the term self-control encompasses all strategies for dealing with willpower limitations, including situational or environmental modifications. These situational strategies to avoid temptation are typically classified under self-control rather than willpower because they rely on external mechanisms rather

than internal mental effort (Fujita, 2011; Hofman & Kotabe, 2012). Self-control is hence a broader term that includes both willpower and external mechanisms to resist temptation. In this chapter, we use willpower when concerned with the fundamental limitation or the internal struggle to cope with temptation and self-control when concerned with all strategies for coping with temptation. When discussing other authors' work, we adhere to their specific terminology to represent their ideas and arguments as accurately as possible.

From a rational choice theory perspective, the need for self-control poses a puzzle because why would rational agents need self-control? Why would someone act against their own best interest? Yet, as behavioural economists have stressed, agents frequently do report acting against their own best interest and against their self-formulated preferences (Thaler, 2015; see chapter 4). Self-control is thus necessary because agents experience an internal conflict; part of them prefers x, while another part favours its negation. This conflict can be understood as a tension between two conceptual selves, with one associated with system 1 thinking and short-term preferences, and the other with system 2 thinking and long-term preferences. We use the terms short-term self and long-term self metaphorically, to capture the different cognitive modes and temporal orientations, and to align with the literature on self-control, rather than claiming any ontological distinction between them.

The literature on behavioural economics, and (the ethics of) nudging, usually conceptualize this internal conflict as acting against one's higher-order preferences (Thaler, 2015). Jolls, Sunstein, and Thaler define bounded willpower as: "the fact that human beings often take actions that they know to be in conflict with their own long-term interests" (Jolls et al., 1998, p. 1479). These accounts stress that willpower is necessary because agents otherwise act against their “real” preferences and thereby hurt their own welfare (see chapter 4). The dominant conception of willpower in behavioural economics holds that

willpower is the active and effortful suppression of short-term preferences to act in accordance with long-term preferences.

Inspired by Ainslie's account (2021), we argue that the dominant conception of willpower is not completely false but is overly simplistic because it favours system 2 over system 1 and implies that long-term preferences are necessarily correct and complete. We argue that this premise precludes a meaningful analysis of how digital nudges hurt willpower and welfare as it emphasizes only effortful willpower strategies while neglecting (strategies to develop) effortless willpower while the latter are more promising.

In contrast to the dominant view, we conceptualize bounded willpower as "the widespread violation of temporal consistency" (Ainslie, 2006, p. 4). This definition likewise highlights that willpower concerns a conflict between short- and long-term preferences, often mapped onto System 1 and System 2. However, it does not assume that long-term, reflective preferences are automatically authoritative.

A common objection to this more balanced view is that critical reflection appears to belong exclusively to System 2. If so, willpower would seem to consist simply in reflective override: System 2 correcting or suppressing System 1. Yet this view is too narrow. While reflective evaluation originates in System 2, willpower cannot be reduced to momentary acts of effortful suppression. Rather, durable willpower requires that reflective commitments and automatic motivational processes become increasingly aligned.

Here Frankfurt's notion of higher-order endorsement is helpful. On his account, weakness of will does not merely involve the presence of competing desires, but a failure of wholehearted identification. What strengthens the will is not the brute dominance of a higher-order preference, but the integration of one's motivational structure such that endorsed aims are stably supported by first-order desires. Willpower, in this sense, depends on coherence between levels of motivation, not simply on effortful reflective correction.

Similarly, as Tiberius suggests, the task is not to pit rational capacities against emotions, but to cultivate reflective capacities so that they function together with our moods and desires in guiding us toward our aims. This implies that preference shaping must occur on both sides of the conflict. Reflection can clarify, evaluate, and endorse ends, but the motivational force necessary for sustained action depends on reshaping habits, attention, and affective responses within System 1. As Tiberius (2008) formulates it: the "strategy is to think about how to train the rational and reflective capacities we actually have so that they can function together with our emotions, moods, and desires to get us somewhere we'd like to be" (p. 7).

This dynamic is captured by Haidt's rider-and-elephant metaphor, which we discussed in the introduction of this dissertation. The rider (reflective reasoning) may set direction, but lasting control depends on training the elephant (automatic motivational processes). Riders who constantly pull against an untrained elephant will quickly exhaust themselves. Likewise, a conception of willpower that relies solely on effortful override misunderstands the phenomenon. Stable willpower is achieved when reflective commitments and automatic dispositions are mutually reinforcing rather than chronically opposed.

A more balanced view of willpower therefore requires preference shaping at both levels. It is not merely about choosing the long-term over the short-term in isolated moments of temptation, but about structuring one's hierarchy of preferences so that endorsed aims are supported rather than undermined by one's automatic tendencies. Only then does willpower become robust rather than fragile.

Importantly, willpower is only concerned with conflicts between preferences that are recognized by an agent. As explained in the previous chapter, goal-conflict recognition and goal-conflict identification are concerned with understanding one's preferences (Myrseth & Fishbach, 2009). Here we are

concerned with acting on that understanding, which Myrseth and Fishbach (2009) refer to as goal-conflict resolution.

After broadening the definition of willpower as commonly used in the literature on behavioural economics, we analyse how digital nudges hurt willpower in two ways, and how this affects proficient shaping of preferences and welfare. First, digital nudges amplify the need for self-control. Second, they diminish the effectiveness of self-control strategies. We conclude with recommendations on how educators and educational policymakers can teach students to protect themselves against (the adverse effects of) digital nudges on willpower and thereby on welfare.

7.2 Self-control in the behavioural economic literature

Rational choice theory posits that agents possess complete and consistent preferences. In addition, it is assumed that their actions perfectly reveal those preferences, so that no difference between preferences and choices exists.

Under this assumption, willpower is made superfluous, highlighting the psychological unrealistic nature of the assumption (Cowen, 1991). Willpower is only necessary for someone experiencing the urge to act in conflict with their preferences. For example, for dieters, willpower might be necessary to refuse a tasty-looking chocolate desert. Resisting the urge to indulge might be referred to as an act of willpower, the opposite as weakness of will. It consists in a tension between two preferences that are mutually exclusive, one of which is a preference for something of which the rewards will be experienced on the short-term (eating desert) and one for which the rewards will only be experienced on the long-term (dieting). Many agents exhibit inconsistent preferences over time, which is attributed to a lack of willpower (Thaler, 2015; see chapter 4).

Of the three fundamental limitations, which we defined in chapter 4, George Ainslie views a lack of willpower as the most serious (Ainslie, 2006). The

literature on behavioural economics extensively documented that people frequently act against their own best interests, struggle to act on self-formulated goals, and experience temptations and conflicting desires.

Behavioural economics has attempted to integrate this insight into economic theory. The starting point for all behavioural economic thinking on self-control is recognizing that the term presupposes a conflict within the self (Thaler & Shefrin, 1981; Duckworth, 2011). One part of the self wants to indulge while another part wants to resist. As Angela Duckworth, one of the leading authors on the psychology of self-control formulates it: the notion of self-control "presumes an internal conflict between mutually exclusive responses" (Duckworth, 2011, p. 2639). In addition, the notion of conflict of the self implies the existence of two entities within the self. Richard Thaler (2015) quotes David McIntosh: "[t]he idea of self-control is paradoxical unless it is assumed that the psyche contains more than one energy system, and that these energy systems have some degree of independence from each other" (McIntosh, 1969, p. 122). These two entities are commonly referred to as two "selves" (Schelling, 1984, p. 8) in the literature on behavioural (welfare) economics, although how authors view the ontological status of these selves varies (Schelling, 1984).

7.2.1 Time preferences

The conflict between the two selves is often depicted as an intertemporal conflict, as a conflict between short-term preferences and long-term preferences. In many behavioural economic experiments concerning willpower participants must choose between two options that pay off over different time horizons, with one option offering a smaller reward sooner ("smaller sooner (SS)") and the other a larger reward later ("larger later (LL)") (Ainslie, 2021, p. 6). These experiments are designed to reveal through intertemporal choices how agents value time. Many dietary and financial decisions can be described in this way.

For example, should one buy a leisurely good, or save for their pension? Should one eat a delicious desert, or diet?

Behavioural experiments show that people choose very differently when choosing between a smaller reward 30 days from now and a bigger reward 31 days from now versus between now and tomorrow. In other words, they value the same amount of time differently when it is closer. This means that many agents discount the future hyperbolically, causing them to exhibit time-inconsistent preferences (Hoch & Loewenstein, 1991; see chapter 4). As a result, many preferences for options with smaller rewards sooner are temporary, illustrating the inconsistency of intertemporal preferences (Ainslie, 2021). This temporary state of wanting to indulge by choosing options with smaller rewards sooner is therefore referred to as the “hot state”, as opposed from the “cool state” (Ainslie, 2021, p. 9).

7.2.2 Two selves in conflict

Self-control conflicts can be thought of as interpersonal conflicts. As conflicts require two parties, self-control can be seen as a battle for control. As a result, self-control has been modelled by positing that agents have two selves with "two separate sets of preferences" (Thaler & Shefrin, 1981, p. 2). This framework enables analysis of self-control through an agency relationship perspective. For instance, Thaler & Shefrin (1981) apply the principal-agent model to internal conflict, depicting a strategic rule-setter and an executive who adheres to these rules.

A common depiction contrasts one self as impulsive, focused on short-term preferences and irrational, while the other is deliberate, focused on long-term preferences and rational (Cowen, 1991). This characterization remains prevalent, particularly in the literature on nudging that emerged after 2008. The two selves have been assigned various names depending on the author's perspective. Thaler & Shefrin use "doer" and "planner" (1981, p. 2), a

terminology also adopted in *Nudge* (Thaler & Sunstein, 2008). Cowen refers to them as the "impulsive self" and the "rule-oriented self" (1991, p. 361). In contemporary behavioural (welfare) economics, the beta-delta model introduced by Laibson (1997) dominates intertemporal preference analysis (Laibson & Maxted, 2023). This model captures present-biased preferences by applying additional discounting to future periods, mathematically representing the tension between immediate and delayed gratification. Within this framework, "present self" and "future self" or "future selves" have become the standard terminology (Bleichrodt et al., 2022, p. 6326).

The two selves described here seem similar to system 1- and system 2 thinking as described in dual process theory. As Thaler states: "it is reasonable to think of the planner as the slow, reflective, contemplative system 2 while the doers are the fast, impulsive, intuitive system 1." (Thaler, 2015, p. 109). Similarly, the doer and planner described by Thaler and Shefrin (1981) and the impulsive self and the rule-oriented self of Cowen (1991) could be interpreted as corresponding to system 1 and system 2, even though that terminology that was only invented after their writings (Stanovich (1999). Hofmann et al. (2009) also analyse self-control through the lens of dual process theory.

7.2.3 Suppression

In this view, self-control conflicts centre around resisting or suppressing the impulses of system 1 and acting on the long-term commitments of system 2. The main strategy to achieve this, is to keep temptations from one's attentional awareness, or if they reach the attentional awareness, preventing oneself from considering or evaluating the temptation. Ainslie calls this strategy for achieving willpower "suppression" (Ainslie, 2021, p. 13). "Suppression is a way of gating out alternatives to a current intention while ignoring their possible value" (2021, p. 5).

Suppression is highly effortful as it relies on an agent's ability to steer their attention away from where it naturally flows. We have highlighted in previous chapters how system 1 is automatic and how its biggest strength is low cognitive expenditure. When agents continuously need to consciously intervene in their natural tendencies, this is highly effortful. Therefore, Ainslie argues suppression is "necessarily unstable" (Ainslie, 2021, p. 5).

Many authors in the behavioural economic literature equate self-control with suppression and therefore argue self-control is necessarily limited. They depict self-control as a muscle, which fatigues after use. Roy Baumeister, who created the influential, yet controversial ego-depletion theory of willpower is the most important proponent of this theory (Baumeister et al., 2008; 2024).

Suppression and ego-depletion have been linked to dual process theory. According to Baumeister et al. (2008) ego depletion leads to deterioration of system 2 processes but not system 1 processes. This fits with the view that system 1 is in default mode and is more likely to act on lower-level preferences, while its system 2's task to intervene. They portray intervention as a costly process that diminishes in effectiveness when the agent is tired or "ego depleted". Similarly, Levy argues: "that the resemblance between ego depletion and the effects of cognitive load is not a coincidence: self-control is a system 2 process, and its loss switches us to system 1" (Levy, 2011, p. 7). Gillebaart and de Ridder also recognize that “although it is mostly implied rather than explicitly stated, a common way to consider self-control in the framework of dual-systems theories is to regard self-control as part of system 2: the ‘cool’ system that is able to override impulsive output from system 1” (2017, p. 38). This is in line with Thaler’s (2015) statement that the planner can be thought of as system 2, and the doer as system 1. Many authors in the literature on behavioural (welfare) economics have (implicitly) endorsed this view, conceptualizing self-control as the ability to stop system 1 from automatically reacting to its natural impulses by intervention of system 2, a process that is necessarily limited.

7.2.4 Situational self-control strategies

Alternatively, agents could exercise self-control through "situation modification strategies" or “situational strategies” (Duckworth et al., 2016, p. 8) such as avoiding temptations by keeping a physical distance, asking others not to offer certain temptations, or pre-committing oneself. Situational or external control strategies might be contrasted with "intrapsychic" (Duckworth et al., 2016, p. 7), or internal control strategies. Hofmann & Kotabe (2012) denote a similar distinction with preventive and interventive self-control. The crucial difference between the two forms of self-control is that the first strategies are used before craving feelings of temptation arise, while the latter are used when these feelings are present (Ainslie, 2021).

Duckworth et al. (2016) note that situational self-control strategies may be underappreciated and suggests this may be because they prevent the effortful resistance to indulging that many people associate with self-control. In addition, situational control strategies have the potential to be highly effective without ever being noticed by the agent.

Situational strategies for self-control and especially pre-commitment strategies can be considered a tactic by the long-term self to make a final decision before the hyperbolic discounting takes place and the perceived value of a smaller reward surpasses that of a larger reward which will be obtained later. The long-term self hereby circumvents the short-term self and presents it with a fait accompli. In some cases, this could be noticed by the short-term self at the moment the pre-commitment becomes visible. For example, agents who try to avoid alcohol but go to a restaurant could promise their friends to be the designated driver, so they are pre-committed to remaining sober. When the friends order their drink, the driver will notice and feel their pre-commitment taking effect. However, some pre-commitments may not have the same moment

of notice. For example, when someone asks their spouse no longer to buy alcohol, they may never notice the effect of their pre-commitment.

7.3 Critiquing the prioritization of system 2 in willpower discussions

Thaler & Shefrin's (1981) Beta-Delta model, which is influential in behavioural economics (Gabaix & Laibson, 2017; Beshears et al., 2016; Bleichrodt et al., 2022; Laibson & Maxted, 2023), and the literature on (the ethics of) nudging (Thaler & Sunstein, 2008; Thaler, 2015) all emphasize the importance of the long-term or system 2 self. In their view, agent welfare "depends upon the ability of the rule-oriented [system 2] self to win the battle for [self-control]" (Cowen, 1991, p. 365). From this perspective on self-control, which we call the naive view, self-control is often framed as the ability to resist system 1 urges, which are seen primarily as obstacles to long-term welfare. By prioritizing system 2 over system 1, the model suggests that self-control is about ensuring that system 2 thinking prevails over impulsive urges. While this view is not entirely false, it is overly simplistic because it assumes that favouring the long-run self always leads to greater welfare for the agent. As we discussed in the previous chapter, while overreliance on system 1 can indeed cause an overemphasis of short-term preferences, system 1 also plays a valuable role in forming stable habits and maintaining self-control. On the other hand, exclusive reliance on system 2 may lead to overthinking and a detachment from the immediate, context-sensitive realities that system 1 helps address.

7.3.1 Favouring system 2 presupposes complete and consistent preferences

In Thaler & Shefrin's (1981) model of the planner and the doer, a skilful doer is supposed to execute what is decided by the planner. This assumption is

made explicit by their choice of comparing the relationship between planner and doer to the principal-agent model. In this economic framework, principals are modelled to have an interest that must be achieved. Principals alone are responsible for deciding what must happen and setting up the rules for how this must happen. Agents are assumed to have no stake in the result, other than honouring his contract with the principals. Choosing the principal-agent framework for a comparison with system 1 and system 2 creates an image of a deciding system 2 that must keep system 1 under control.

This view of system 1 and system 2 is consistent with the welfare conception of behavioural economics which we critiqued in chapter 4. There, we argued that behavioural economics' laundered preferences are in essence an attempt to model agents as possessing complete and consistent preferences. Agents only err because of psychological forces. We referred to this critique as "an inner rational agent is trapped in an outer psychological shell" (Infante et al., 2015, p. 1).

The willpower discussion is a subset of this model focused on time preferences. System 2 is modelled as having correct time preferences, which are then troubled by the outer psychological shell, or the present bias (Beta) that interferes and leads to inconsistencies which must be banished. Thaler shares this view, when he states that that the "crucial advantage" of the beta-delta model is its "mathematical simplicity" (Thaler, 2015, p. 110). According to him, the model is "the smallest possible modification of Samuelson’s basic model [of revealed preferences] that can capture the essential aspects of self- control" (Thaler, 2015, p. 110).

Optimizing for mathematical simplicity and proximity to Samuelson's model is problematic because mathematical models require complete preferences, even when none exist. As argued in chapter 4, this can be seen as a problematic attempt to continue orthodox economics, only by introducing disruptive psychological forces that divert agents from their otherwise rational (as defined by rational choice theory) course.

7.3.2 The naïve view on self-control neglects system 1's ability to engage in strategic behaviour

Thaler & Shefrin's principal-agent model depicts the two selves as two agents, engaged in a strategic game. However, only the principal, or system 2, is capable of strategic behaviour while the agent, system 1, is not. This contrasts with Cowen's (1991) view, who argues that system 1 is perfectly capable of strategic behaviour. Ainslie (2021) also views system 1 as capable of strategic actions and explains how this feature is crucial for understanding the self-control strategy he refers to as "resolve" (p. 5). We discuss this argument in depth in the next section.

7.3.3 Overreliance on system 2 is harmful to self-control and welfare

While many authors in the literature on the ethics of nudging view system 1 as the source of error and system 2 the source of rational preferences, "[t]here is no prima facie reason for believing that victory for the rule-oriented self [system 2] is desirable" (Cowen 1991, p. 365). In fact, "excessive focus [...] on the power of conscious thought and the benefits of willpower and self-control [in some areas of life is] profoundly counterproductive" (Slingerland, 2014, p. 13). As Cowen notes: "too many victories for the rule-oriented [long-term] self can be injurious to mental health. A person who continually thwarts the desires of his impulsive self may become frustrated and overly rigid and lose his capacity for spontaneity" (Cowen, 1991, p. 365). Likewise, Ainslie (2001) argues that "nothing fails like success" (p. 141) because an excessively strong will undermines appetite.

Stated even starker, long-term preferences can be reason for failures of self-control. While initially counterintuitive, Arpaly discusses "cases in which an

agent is more rational for acting against her best judgment than she would have been otherwise" (Arpaly, 2000, p. 504). She discusses Emily whose long-term preference and reflective judgment has always been that she should pursue a chemistry PhD. But when she takes steps towards this goal, such as completing her studies, she experiences increasing feelings of unrest, sadness and anxiety. Performing tasks for her studies requires an increasing amount of self-control. However, Arpaly emphasizes, "when she deliberates, she concludes that her feelings are senseless and groundless" (p. 504). When one day she impulsively quits the program, this must be considered a lapse of self-control.

This example function to illustrate that in some situations, problems of self-control are not created by an overly impulsive short-term preferences but by overly controlling and demanding long-term preferences. One could argue that Emily does not actually prefer to pursue a PhD in chemistry but that is no different from dieters who prefer not to prefer chocolate but in reality, do.

If we applied the naive view to digital nudges, the argument in this chapter would be straightforward: digital nudges would be seen as leading to an overuse of system 1, making people more susceptible to temptation and, consequently, to welfare-reducing choices. The solution to achieving self-control would then be to either circumvent system 1 with situational control strategies or suppress system 1 urges. While these are valid self-control strategies, this perspective neglects the most promising one—resolve, which fosters habit formation. It also obscures the importance of shaping preferences by implying that system 2 contains fully formed, complete preferences.

This bias in favour of system 2 is deeply embedded in both behavioural economics and the literature on (the ethics of) nudging. As we argued in Chapter 4, it reflects a broader issue in behavioural economic welfare thinking: the conception of the rational agent in a psychological shell. Addressing this issue is crucial because it prevents a meaningful analysis of the role system 1 plays in self-control and, in doing so, overlooks how resolve and habit formation enable

agents to engage their full cognitive capacities to shape and fulfil their preferences proficiently.

7.4 Self-control based on resolve

To formulate an account of self-control that has a more balanced view of the two selves, we can take inspiration from Schelling (1984; 1985), Cowen (1991), and Ainslie (2006; 2021). These authors agree with the naïve view that self-control is best conceptualized as an agent in conflict, and that the conflict can be described as multiple selves competing for control. However, they have a different view on the strategic capabilities of system 1 and the role of automatic and habitual behaviour in self-control. This view on self-control is based on what Ainslie calls "resolve" (20021, p. 5), an effortless strategy for self-control which emphasizes adopting a different perspective on conflicted choices, broadens the naïve view on willpower. In addition, we seek to emphasize more starkly how self-control is relevant for welfare and preference shaping.

7.4.1 Resolve

Ainslie (2021) emphasizes resolve as an important strategy for self-control. Resolve is the skill to recognize one's current choice as a signal for future choices. Reframing choices as test cases for future choices makes the value of maintaining self-control clearer. In addition, the motivational impact of reframing a choice as a test case often outweighs the immediate stakes of the decision and can be crucial to maintaining one's commitments. For Ainslie, resolve is different from suppression in that it is less direct in dealing with the temptation, it takes a longer time horizon, and it is a more stable form of self-control. We discuss resolve as the third strategy for self-control, in addition to suppression and situational control.

Ainslie's conception of self-control can account for resolve because it recognizes that the short-term self is capable of strategic behaviour, creating a different bargaining dynamic than in models where the short-term self is uncapable of strategic behaviour. He hence rejects the principal-agent model and instead model the relationship between the two selves as a (roughly) equal relationship between two agents who both act strategically to reach their goals.

In a strategic bargaining process, the long-term self can improve its position by pre-committing. For example, if an agent has a long-term preference to complete a difficult task such as writing a philosophy paper, but they know they might fall for scrolling on social media instead, they can install a social media blocker on their computer. This inhibits the short-term self from winning the battle for self-control, committing the agent to acting on their long-term preferences.

Cowen (1991) notes that while the short-term self is unable to reversely pre-commit to spontaneous action, it can increase the cost of pre-commitment by "retaliating" (p. 364) against overly strict pre-commitments or by rebelling more intensely when the opportunity arises. Such actions can be described as counter moves in a repeated strategic game between two players (Ainslie, 2021). For example, after spending the entire day working diligently on the philosophy paper, the short-term self can retaliate when the constraint of the blocker is lifting. It could do so by suggesting scrolling the entire evening on social media, or to consume especially shallow content (see chapter 5). These would be sound strategies from a game-theoretical perspective on two agents competing for control but is also psychologically realistic.

Ainslie conceptualizes the bargaining between the two selves as "an inter-temporal variant of the repeated prisoner's dilemma" (Ainslie, 2021, p. 20). This formulation suggests that both selves are capable of strategic behaviour. In Ainslie's framework, the conflict resembles a prisoner's dilemma because each self faces the choice to either cooperate or defect in relation to the other. He

argues: "conflicts among individuals can be solved by finding clear, albeit often tacit, criteria for what constitutes cooperation or defection, as long as mutual cooperation will benefit each player more than mutual defection will." (Ainslie, 2006, p. 15).

Ainslie frames every decision between an option with a short-term reward and one with a long-term reward—where the agent has previously intended to choose the long-term reward—as a "test case" (Ainslie, 2021, p. 20). On the surface, there may appear to be little reason to avoid smoking a single cigarette or spending one minute on social media, as the immediate negative effects of such actions are negligible. However, Ainslie argues that when these actions are framed as test cases for future behaviour, their harm lies not in their direct consequences but in the damage they inflict on the credibility of one's diet, intentions, personal rules, or character (Ainslie, 2021). Recognizing the motivational impact of interpreting one's current choice as a signal for future choices often outweighs the immediate stakes of the decision and can be crucial to maintaining one's commitments. Ainslie refers to the skill to recognize this as "resolve" (Ainslie, 2021, p. 5).

Ainslie argues that repeated successful resolve “may result in reliable preference for these outcomes" (Ainslie, 2021, p. 2). Once a preference is learned and is stable, choosing in line with this preference "is experienced as an effortless habit" (Ainslie, 2021, p. 2). It becomes automatic as it is no longer exclusively the result of effortful reflection, based in system 2, but an integrated part of system 1. "Essentially, habit reflects bargains between impulses and resolutions that are no longer contested" (Ainslie, 2021, p. 35).

Slingerland makes a similar argument without using Ainslie's concept of resolve. He argues that “the conscious mind can acquire new, desirable goals and then download them onto the unconscious self, where they can then be turned into habits and implemented without the need for constant monitoring” (Slingerland, 2014, p. 62). So, learning to recognize choices as test cases for

future decisions can result in stable preferences at the habitual level. This is the successful implementation of resolve. Agents that achieve this, act in accordance with their higher-level preferences effortlessly and automatically.

Crucially, the naive view which relies solely on suppression neglects strategies for self-control that function without continuous effort, namely resolve and the resulting habit-formation (Ainslie 2021). Recognition of system 1's positive role in willpower represents a significant departure from the naive view and suppression-based models, offering a more nuanced understanding of self-control mechanisms. Successful self-control is not necessarily about perpetually resisting or overriding system 1 urges. Rather, system 1's automaticity can positively contribute to self-control. This perspective also challenges the view that system 1 can be equated with short-term preferences. While it is true that overuse of system 1 causes an overemphasis of short-term preferences, system 1 can also contribute to attaining long-term preferences by building habits.

7.4.2 Alignment and preference shaping

When agents form durable habits which are in line with their long-term preferences, their short- and long-term preferences (system 1 and system 2) are in alignment. Alignment means the intertemporal bargain has been settled, both sides agree. It manifests in what Frankfurt describes as "coherence" or "harmony" (1988, p. 164), where an individual's various motivational forces work in concert rather than opposition. It is the opposite of Luc Boven's "fragmented self" (2009, p. 8). Such alignment causes reduced cognitive load and increases welfare as individuals no longer experience the internal conflict characteristic of self-control challenges. So, habits that are the result of resolve offer the best of both worlds: alignment between lower- and higher-order preferences, and spontaneity and effortless action.

However, learning habits is not easy, and they cannot be learned at will. One cannot simply decide to have new habits. There is a substantial popular self-help literature on learning habits (cf. Fogg, 2020; Clear, 2018; Duhigg, 2013). This literature emphasizes the importance of learning one new habit at the same time, rather than trying to create many new habits at once. Moreover, when learning new habits, it must be taken into account that system 1 is responsible for automatic and habitual behaviour and is dominated by evolutionary preferences which are "gene-installed" (Stanovich, 2004, p. 182). System 2, while more flexible in its preferences, is still influenced by gene-installed goals, and importantly is influenced by what Stanovich calls “meme-installed” preferences. These are preferences that are the result of culture and the environment. While habits can be modified, this process requires acknowledging both our evolutionary predispositions and the cultural influences that shape preferences, rather than relying solely on willpower or disregarding the contextual environment.

Creating habits and influencing automatic thinking processes is a complex exercise. The unique strength of system 1 is that it can process more information than humans could ever consciously reflect on. So, rather than to try to control it, agents should create the conditions to let the automatic processes effectively coordinate. Cowen (1991) draws a parallel between thinking processes and economies. He points to the failure of state-lead economies that cannot centralize sufficient information for effective decision-making. Conversely, market economies are not ruled centrally but have an effective coordination mechanism that operates without anyone necessarily understanding it. In a similar vein, self-control requires the “unleashing of forces in such a way as to create a complex but coordinated process of personality growth" (p. 373). According to Cowen, this coordination cannot be achieved if one self attempts to control the other completely. Treating the reflective system as a central planner that continuously overrides automatic processes neglects the distinctive strengths of System 1,

including its capacity to process vast amounts of information and to generate motivational force. As argued above, willpower is not exhausted by momentary acts of suppression. Rather, it depends on the gradual alignment of reflectively endorsed commitments with automatic tendencies. Equality between the two selves should therefore be understood as complementarity: reflection on motivational forces and first-order desires strengthens them over time.

When such alignment is achieved, self-control no longer relies primarily on effortful resistance. Instead, action flows from a relatively stable integration of higher- and lower-order preferences. This integration constitutes the basis of effective self-control and ultimately supports welfare, because agents are less frequently divided against themselves and better able to act in accordance with their endorsed preferences over time. As Schelling states: the “phenomenon of rational strategic interaction among alternating preferences is a significant part of most people's decisions and welfare” (Schelling, 1984, p. 5). This discussion shows that self-control is not simply resisting every impulse or short-term preference. Rather, it is the psychological skill to manage impulses in a way that achieves continuous alignment between lower- and higher-order preferences. Agents can use situational control, suppression or resolve as strategies against unwanted impulses, where successful resolve can result in stable habits. By iterating spontaneous action and reflection, agents can learn to act and then reflect on their actions and formulate new rules to adjust their natural tendencies aligned with their reflectively endorsed second-order preferences. This is ultimately what preference shaping is. This process must occur with regard for the short-term as well as the long-term, and with respect for our natural characteristics and capabilities.

7.5 How digital nudges undermine self-control strategies

Digital nudges reduce the effectiveness of self-control strategies, and thereby welfare. They do so by negatively impacting all three strategies for self-control, namely situational control, suppression and resolve, the latter which should result in habit. In this section we analyse how digital nudges negatively impact all strategies. In the recommendation section we analyse how agents can nonetheless improve their strategies for self-control when confronted with digital nudges.

7.5.1 Situational control

First, digital nudges diminish the effectiveness of situational self-control strategies because these strategies rely on the agent's ability to control their environment, while digital nudges are intertwined with the environment. As Reijula & Hertwig state: “[i]t is hard to think of a modern environment with an architecture more skilfully designed to capture and command control over human behaviours than the online world.” (2022, p. 131). In addition, digital products are characterized by the opposite: control over the environment for the designer rather than for the user.

It is hard for digital users to recognize, anticipate, and avoid digital nudges. Reconsider the example of dieters, who can learn to recognize the moments they are most likely to lapse. For example, there are many occasions where it is culturally appropriate to serve unhealthy food. When visiting a birthday, no one should be surprised if cake is served. Subsequently, dieters can avoid particularly dangerous situations and environments to avoid temptation (without suggesting this would be easy). While cravings for digital media consumption could likewise be predicted and anticipated to an extent, their lure is not as closely connected to time and place. Instead, its ubiquity ensures that

moments to lapse are more frequent and harder to anticipate. Moreover, digital nudges in the shape of push notifications and pop-up messages can initiate the experience autonomously, making it harder for users to avoid their temptation by leveraging situational self-control strategies. Finally, as smartphones are general purpose machines, they prohibit a clear separation of environments. For example, many people use their smartphone as a bedside or wake-up alarm, making it the last thing they see before falling asleep and the first thing they check upon waking. Similarly, Schuster et al. (2023) found that adults who "self-identify as smartphone over-users" report having trouble managing their smartphone use due to an unclear boundary between work and personal use of devices. They also report experiencing pressure to respond quickly to work emails, providing them with an incentive to keep their smartphone close. This combination of features makes digital nudges difficult to recognize and anticipate, complicating the use of situational self-control strategies.

In addition, the digital environments are harder to control. Revisiting the example of dieters, a common saying is: ‘don't buy it, and you won't eat it’. If unhealthy foods are not in your home, it is easier to maintain a healthy diet. Dieters can do this because they are free to arrange their homes as they want, including the kitchen, cupboards and pantry. Doing so helps them exercise control over their diet-related behaviours and thereby leveraging self-control. Conversely, digital media users have limited control over their digital "pantry" because digital designers have a large degree of control over both the device and the applications and websites on the device. While we discussed the ways in which digital designers control devices in chapter 3, one important example is this context is the use of Android phones. The Android operating system is created and owned by Alphabet and operates on approximately 70% of all phones globally (Statista, 2025). Android phones pre-install and digitally nudge consumers to use other Alphabet services such as Google search and YouTube, where YouTube is a huge source of distraction and additional digital users for

many people. In addition, some phone manufacturers like Samsung also pre-install or digitally nudge users towards social media apps such as Facebook, TikTok, and Snapchat. Moreover, the manufacturer installed apps can be deleted but the Android apps cannot be deleted, only "turned off" which means they can always come back with one click and with the friction of installation time removed. So, for digital users, it is highly challenging to control their environments. Of course, heavy addicts could choose to abandon their smartphones completely but to extent the dieter analogy, doing so would be like removing the cupboards and kitchen from your house altogether. Smartphones and other digital devices have become socially expected and are necessary to function in society. In addition, many people receive a preconfigured device from their work which they cannot choose.

The combination of the ubiquity and availability of smartphones with the fact that the trigger to begin smartphone use makes it difficult to use situational self-control strategies to limit excessive digital media usage. However, this is not to suggest that nothing can be done against digital nudges, as we discuss in the recommendation section.

7.5.2 Suppression

Second, even without considering the effect of digital nudges, suppression is not the most effective nor most durable strategy for self-control, as it is "necessarily unstable" (Ainslie, 2021, p. 5). Suppression, the act of temporarily blocking thoughts about tempting alternatives, pretending they have no value, or diverting attention, is a temporary solution to self-control. Doing so cannot be maintained indefinitely, especially because doing so would require enormous amounts of cognitive capacity and would be highly effortful. Formulated in terms of Baumeister's theory (Baumeister et al., 2024), continuously relying on suppression would be quickly ego-depleting. Nevertheless, given its prominence in the literature, it is worth emphasizing that digital nudges significantly

undermine the effectiveness of this approach. The sheer ubiquity, overwhelming volume, and potency of digital nudges (as previously argued) make resisting them through suppression unsustainably demanding in terms of the cognitive effort required.

In many examples of temptation, agents prefer a larger reward in the distant future more than a smaller reward in the near future. However, when time progresses, and both rewards come closer, the smaller reward is discounted faster, which causes the option with the smaller reward to (temporarily) become the preferred option. An often-mentioned example is of dieters who prefer to decline sweets, until the sweets become visible. When they indulge, regret may follow soon after. This model is partially applicable to digital nudges because many digital nudges can be described as aiming to make the desirable characteristics of an option more salient and thereby more strongly felt. As digital nudges target shallow cognitive processes, they are particularly successful in making the desirable features of options with smaller sooner rewards more strongly felt. In other words: digital nudges make the cake seem even sweeter and more accessible. Digital nudges create this effect due to several of their characteristics.

The first characteristic is that digital nudges increase the need for self-control because they are ubiquitously present. In previous chapters we showed data on how much time agents spend in digital environments. In addition, we highlighted the extensive control digital nudgers have over these environments. This combination causes most people to be confronted with large numbers of digital nudges every day. We have called this a sea of nudges. Moreover, we argued that many digital nudges intent to maximize user screen time and capture as much time and attention as possible, due to the financial incentives.

These characteristics of digital nudges increase the need for self-control of digital media consumers. Imagine that to respect their diet, dieters need to keep their level of self-control above a critical threshold to avoid breaking their

dietary commitment. There may be many moments during the day that the required level of self-control falls below the critical threshold, for example during an annoyingly long work meeting late in the day, during a tedious car ride, or after exercising. However, these dieters might still adhere to their diet because at the time of the shortage of self-control, there may be nothing to indulge with available. The time needed to find a sweet may be enough to restore sufficient self-control to resist temptation.

In contrasts, individuals struggling to resist overconsumption of (shallow) digital information are in presence of their object of temptation almost permanently. Many people carry their smartphones everywhere and all the time. As Hofmann et al. formulate it: the "constant availability of media content has greatly expanded our immediate access to a plethora of media gratifications as we now have a virtually unlimited arsenal of media resources at our disposal at any time and in any place" (2016, p. 2). This means that digital distractions are instant and constant, leaving little opportunity for self-control to recover when it inevitably falls below the critical threshold occasionally.

The second characteristic is that digital nudges increase (the salience of) temptations, they particularly amplify the urge to seek instant gratification by consuming shallow digital information. In addition to the ubiquitous availability of digital media, their availability is instantaneous. For people struggling with overconsumption of digital media, it takes a second to take reach for their phones in their pocket and open any social media app. Once open, the gratification of novel information in nearly instant.

As we argued in previous chapters, the types of information that perform well on digital platforms are those exploiting human biases, such as the biases for social and negative information. This is exacerbated by personalization, allowing digital media users to be served the type of information that is uniquely tempting to them. Digital nudges and online platforms have no intention to serve users any particular type of content. Their only objective, which is codified in the

algorithms, is maximizing screen time. As a result, content that gains traction and reaches large audiences is shaped by an evolutionary algorithm that favours the most irresistible and attention-grabbing information.

This makes digital nudges uniquely challenging to maintaining self-control. As Saghai (2013) argues: "[w]hen self-control is at stake, influences reinforcing our natural propensities are more likely to end up substantially controlling our [behaviour]" (Saghai, 2013, p. 491). The personalized algorithms determining which information is served to digital users optimizes for whatever the natural propensities, aligned with system 1, of a unique user are. This makes every algorithmically determined display of information optimized for challenging the self-control of its user.

This challenge to self-control can be better understood by examining the mechanisms that drive digital engagement. Behavioural scientists have identified key factors that determine whether someone will act on an impulse. One influential model is BJ Fogg's framework of behaviour as a function of motivation, ability, and a prompt or trigger (Fogg, 2020). Formulated simply, for an agent to perform a behaviour, they need to be able to do it, want to do it, and be triggered to do it. Triggering can be as subtle as seeing something that causes the agent to think about the behaviour.

The triggers driving user engagement in digital products are often rooted in negative psychological states. Nir Eyal explains in *Hooked* (2014) how "internal triggers" (p. 47) such as feelings of boredom, loneliness, or insecurity, are key motivators for behaviour. These emotional discomforts create a psychological tension that users seek to resolve by turning to specific digital platforms. Eyal highlights that this connection between negative feelings and digital products was intentionally designed by digital designers. For instance, a sense of boredom might prompt someone to scroll through social media, struggling at a difficult work project may lead to seeking distraction through entertainment platforms or casual games, while feelings of inadequacy might

lead to engaging with fitness apps or self-improvement tools. Digital media consumption is hence used to numb negative feelings. They offer the opportunity to escape negativity. Indeed, there is empirical evidence that digital media use causes unwanted distractions and delay (Hofmann et al., 2016; David et al., 2015; Rosen et al., 2013).

Eyal's analysis suggests that successful digital products do not merely respond to user needs but actively leverage negative emotions to pull users to consume digital content. Linking negative feelings to the urge to consume a digital product is a potent form of digital nudging. This mechanism shows how the search for instant gratification is linked to negative feelings, making it hard to resist them and to maintain self-control as those moments of weakness are hard to avoid and catch people at their weakest moments.

Linking digital consumption to negative feelings is detrimental to welfare because the natural propensity of system 1 thinking, which is fast, automatic, and driven by immediate rewards, makes individuals more likely to indulge in instant gratifications. This contrasts sharply with second-order preferences, which often offer greater long-term rewards but require effort, patience, and the ability to endure negative emotions such as frustration or concentration. Activities like learning a new skill or completing a challenging work project demand engagement with these emotions, rather than avoidance. The variable reward schemes we discussed previously are particularly relevant here, as they generate dopamine responses that make users increasingly sensitive to short-term rewards while diminishing their capacity for the delayed gratification necessary for activities with longer time horizon rewards.

As Hofmann et al. (2016) observe, the result of digital nudges is that "users find it increasingly difficult to maintain a balance between the short-term pleasures of media use and the potential costs of media exposure for their long-term goals" (pp 2–3). Digital media, through carefully designed nudges, amplify these temptations, further complicating the use of suppression as a strategy for

self-control. This intentional design makes it harder for users to maintain self-control, causing one of the most common self-control failures that people experience (Hofmann et al., 2016).

7.5.3 Resolve

Third, digital nudges diminish the effectiveness of resolve as a strategy for self-control because resolve relies on an agent's ability to reformulate individual decisions within a broader framework of long-term goals. In contrast, digital nudges excel at capturing attention, exploiting system 1's fast and intuitive decision-making processes, and pulling agents toward immediate gratification. Reformulating a choice in terms of its long-term impact can be understood as a form of self-nudging (Reijula & Hertwig, 2022) by reframing a decision. For example, the decision to play an online game like chess or Candy Crush Saga is no longer framed as a momentary choice, but as a pattern: do I want to be someone who fills quiet moments with online games, or someone who finds contentment in simply being present during quiet moments?

For individuals who have reflectively decided they prefer the latter, this self-nudge aligns with their second-order, long-term desires. However, the challenge lies in maintaining alignment between their reflective, system 2 goals and their immediate, system 1 impulses. When temptation inevitably arises, the short-term self attempts to reframe the situation to justify breaking the rule: *just one game won't hurt,* or *it's Sunday, so the rule doesn't apply*. Ainslie (2021) describes this as the "logic of intertemporal bargaining" (p. 42) between the two selves, where self-control depends on recognizing a long-term rule as dominant and non-negotiable. When the rule is ambiguous or open to exceptions, doubt creeps in, creating tension and mental effort.

Digital nudges exacerbate this process by constantly presenting opportunities to act on short-term impulses, making it easier to justify exceptions and harder to achieve the mental clarity that resolve requires. Many

apps employ tactics such as sending push notifications at carefully optimized moments, displaying red notification badges that trigger curiosity and capture attention, creating a sense of urgency through scarcity cues (*don't miss your streak, limited time offer*), and leveraging social proof (*your friends are watching this livestream*). These tactics effectively "weaponize" system 1—the short-term self—by providing it with persuasive arguments to use in the intertemporal bargaining process.

A second way digital nudges undermine resolve is by making it harder to establish clear rules. As previously discussed, the multipurpose nature of digital devices complicates boundary-setting. For example, many apps use digital nudges in the form of notifications. Notifications may work as genuine reminders for the nudgee, who forgot to undertake an action in the app. However, many notifications are not truly concerned with reminding the nudgee to take some action. Rather, they serve to bring desired behaviour to the forefront of an agent's attention, even though they did not forget about the app. Reminding someone about something they already know does not provide novel information. It does however reduce the (mental) friction to take the action by a form of digital design and is hence a digital nudge.

An agent might set a rule against playing games at certain times, but if they open their smartphone and see a notification from a game, does the rule permit them to read it? Avoiding the notification altogether may prove difficult, especially when games are designed to punish players for ignoring them. In some cases, not reading a notification can harm a player's ability to succeed in the game, creating an all-or-nothing dynamic: either you play with notifications enabled or you don't play at all. By leveraging these tactics, digital nudges not only tempt users but also actively prevent them from creating and adhering to clear rules. This further diminishes the effectiveness of resolve.

So far, we have examined how digital nudges undermine system 2 processes. However, resolve remains the most promising self-control strategy

because personal rules can evolve into habits, transforming deliberate, system 2 actions into automatic, system 1 behaviours. However, digital nudges also disrupt habit formation, making it harder for habits to form and for system 1 to function effectively. In addition, digital nudges not only interfere with this process, they themselves aim to create habits, but often of a counterproductive nature. To understand how digital nudges create habits, we can turn to Nir Eyal's *Hooked* model, which outlines a four-step framework for habit formation in digital products. Eyal (2014) explains that "habit-forming technology" triggers the user's attention (through either internal or external cues); prompt an action with an anticipated reward; deliver a variable reward, triggering dopamine release that suppresses reflective reasoning and amplifies desire; and finally, encourage investment through effort, data, or time (Eyal, 2014, p. 10).

In contrast to self-control strategies that encourage the formation of positive, goal-aligned habits through consistent practice and repetition, digital nudges aim to form habits that reinforce instant gratification and frequent engagement with online products. While self-control strategies are designed to create alignment with long-term preferences, the habits formed by digital nudges serve short-term impulsiveness, creating a tension between these two forms of habit formation. This tension undermines the effectiveness of habit formation as a strategy for self-control, as digital nudges actively cultivate habits that work against the alignment resolve seeks to create.

7.6 Recommendations

To mitigate the harmful effects of digital nudges on self-control and thereby welfare, we propose the following recommendations that are targeted at high schools and educational policy makers and are structured similarly to the previous chapters.

7.6.1 Teach students to recognize the internal conflict between their two selves

By focusing on promising self-control strategies, students can learn to understand and manage their internal bargaining process so they can foster alignment between their two selves and thereby between their preferences. Hereby, we hope to help students to navigate the challenges posed by digital nudges. The first step in an educational program on digital nudges and self-control is to teach students to recognize the internal conflict between their two selves. Crucially, both selves serve a valuable role. A clear understanding of this internal tension lays the groundwork for appreciating why self-control is necessary and how different strategies can support it.

7.6.2 Train students in applying situational control strategies

We have argued that situational self-control strategies are negatively impacted by digital nudges since digital nudges also have the aim of controlling the situation. Moreover, we argued they are hard to anticipate and that digital designers have a great deal of control of the environment and device. However, we also emphasized that situational strategies are still viable.

One effective tactic is to encourage students to be mindful of when and where they engage with digital environments. Setting time- and place-based heuristics—such as avoiding smartphones after 10 p.m. or keeping them stored during meals—can help create useful boundaries. Many schools already implement this by banning smartphones on their premises, a clear example of a "bright line" rule (Ainslie, 2001, p. 94). Encouraging students to routinely place their phones in lockers upon arrival at school fosters both consistency and habit formation.

The same principle can be applied at home. A designated storage space for smartphones, such as a cupboard upon entering the house or a specific place

during meals, helps limit impulsive use. Additionally, reducing reliance on smartphones for essential tasks (e.g., using a separate alarm clock instead of a phone) minimizes incidental use. Another approach is to encourage desktop use over laptops, anchoring digital engagement to a fixed physical space.

The second promising situational tactic is to teach students how to tweak digital environments in their favour. In the previous chapter we recommended tweaking digital environments to introduce friction. Another option is to change settings to avoid temptations. Reijula & Hertwig (2022) for example suggest, changing the homepage of one's web browser from a social media site to a quality news site or setting digital devices to "greyscale to decrease the salience of eye-catching visual notifications" (Reijula & Hertwig, 2022, p. 132). However, especially with social media, tweaking is not the most promising. Social media sites try to push users to install the app by making browser versions of their sites deliberately inaccessible, slow, or otherwise not user-friendly. With Instagram for example, browser users cannot post, create stories or livestream. Direct Message (DM) features and editing tools such as filters and stickers are limited, and the User Experience (UX) is worse. As a result, Instagram effectively forces users to download the app, where customization is more restrained. These all-or-nothing design choices make avoidance or strict usage boundaries more effective than attempts to tweak engagement.

Finally, students should be taught the effectiveness of hard stops. Installing apps that block all distracting apps and websites for a designated period can be highly effective. This tactic could be supplemented with the idea of giving up control to someone else (Schelling, 1984), for example by installing a password for the blocking app that only a friend or family member knows.

7.6.3 Explain the limitations of suppression as a self-control method

Students must learn that suppression, which we defined as actively resisting impulses, is an effortful and unstable strategy for self-control. A first step to understanding why suppression is necessarily unstable as a strategy for self-control and hence for avoiding the temptation of digital nudges is that suppression faces temptation directly, it relies on brute force willpower which is highly effortful. Such effort cannot be sustained indefinitely. In the words of Baumeister et al. (2008), relying on suppression is ego-depleting.

Students should be made aware of the special features of digital nudges that exacerbate the instability of suppression. Digital nudges are intentionally designed to undermine self-control, and their ubiquity, potency, and volume make suppression especially unreliable. Digital nudges increase (the salience of) temptations, they particularly amplify the urge to seek instant gratification by consuming shallow digital information. Moreover, students should understand how the algorithmic personalization of digital nudges progressively undermines suppression strategies. As argued in earlier chapters, digital platforms continuously refine their targeting based on user responses, creating increasingly irresistible temptations tailored to individual vulnerabilities. This dynamic creates a dynamic where suppression capacity remains relatively fixed, while digital nudges become progressively more difficult to resist. Recognizing this fundamental imbalance helps students understand why pure suppression strategies inevitably fail.

Additionally, many digital nudges are linked to what Eyal (2014) calls internal triggers. Eyal's analysis suggests that successful digital products do not merely respond to user needs but actively leverage negative emotions to pull users to consume digital content. These internal triggers make suppression particularly challenging, as they exploit psychological vulnerabilities at moments

when self-control is already compromised. Based on this understanding of suppression's limitations, educators should implement three recommendations to help students manage digital nudges more effectively.

First, students should be encouraged to recognize suppression as a temporary rather than permanent solution. Rather than relying solely on willpower to resist digital temptations, they should view suppression as a short-term bridge to more sustainable strategies. For instance, momentarily resisting the urge to check social media might be acceptable while completing an urgent assignment, but this approach should not be extended indefinitely.

Second, students should be taught to recognize the warning signs of fatigue when using suppression strategies. These may include increasing irritability, difficulty concentrating, or growing preoccupation with the suppressed activity. Baumeister and Tierney (2012) suggest that willpower functions similarly to a muscle that fatigues with overuse. By recognizing these warning signs, students can proactively shift to alternative strategies before complete self-regulatory failure occurs. This metacognitive awareness enables them to understand that diminishing resistance to digital nudges is not necessarily a character flaw but rather an expected outcome of prolonged suppression efforts.

Third, educational programs should emphasize the importance of recovery periods following suppression efforts. To recover, students could engage in physical exercise, mindfulness practices, or engagement with intrinsically rewarding activities, preferably unrelated to digital devices. By incorporating these recovery periods, students can prevent the cumulative depletion that renders suppression increasingly ineffective against digital nudges.

7.6.3 Teach students to resolve temptations through pattern recognition and habit formation

Resolve, as a strategy for self-control, relies on recognizing that individual choices are not isolated but part of broader behavioural patterns. Just as eating one piece of cake does not lead to poor health, but consistently unhealthy eating does, engaging with digital media occasionally is not inherently harmful—but habitual overuse can have significant negative consequences. The key is to shift from evaluating each instance in isolation to thinking in terms of rules or patterns.

For resolve to be effective, students must first develop a clear understanding of the long-term costs of habitual engagement with attention-grabbing algorithms and digital overuse. Without recognizing these dangers, they are more likely to justify each instance of engagement as insignificant rather than seeing the cumulative effect. Education plays a crucial role in fostering this awareness, as schools are well-positioned to teach the broader implications of digital media consumption. By helping students internalize the idea that self-control is about managing patterns rather than resisting each temptation individually, educational programs can support the development of lasting digital habits that protect against the persuasive power of digital nudges.

To translate resolve into stable habit formation, educators should focus on structured routines and reinforcement. Schools can encourage students to set clear, rule-based commitments, such as no phone use during homework time or no social media in the morning. These commitments can then be integrated into daily school activities. Providing students with tools like habit trackers, group challenges, or classroom discussions on digital self-control can help reinforce these rules. In addition, educators should emphasize how these strategies translate to home situations.

Resolve strategies can be effectively combined with situational strategies as situational strategies entail about avoiding digital environments while resolve can set the rule to keep applying the commitment to do so. Educators should emphasize that the goal is to make self-control less about in-the-moment decisions and more about automatic adherence to well-defined personal policies. Over time, this structured approach can help resolve accumulate into durable habits that support self-control to resist the influence of digital nudges to promote welfare.

7.7 Bibliography

Ainslie, G. (2001). *Breakdown of Will*. Cambridge University Press. https://doi.org/10.1017/CBO9781139164191

Ainslie, G. (2006). *A selectionist model of the ego: Implications for self-control*. https://direct.mit.edu/books/edited-volume/chapter-pdf/2318013/9780262283359_caf.pdf

Ainslie, G. (2021). Willpower with and without effort. *Behavioural and Brain Sciences*, *44*, e30. https://doi.org/10.1017/S0140525X20000357

Arpaly, N. (2000). On Acting Rationally against One's Best Judgment. *Ethics*, *110*(3), 488–513. https://doi.org/10.1086/233321

Baumeister, R. F., André, N., Southwick, D. A., & Tice, D. M. (2024). Self-control and limited willpower: Current status of ego depletion theory and research. *Current Opinion in Psychology*, *60*, 101882. https://doi.org/10.1016/j.copsyc.2024.101882

Baumeister, R. F., Sparks, E. A., Stillman, T. F., & Vohs, K. D. (2008). Free will in consumer behaviour: Self-control, ego depletion, and choice. *Journal of Consumer Psychology*, *18*(1), 4–13.

Baumeister, R. F., & Tierney, J. (2012). Willpower: Rediscovering the greatest human strength. Penguin.

Beshears, J., Milkman, K. L., & Schwartzstein, J. (2016). Beyond beta-delta: The emerging economics of personal plans. *American Economic Review*, *106*(5), 430–434.

Bleichrodt, H., Potter Van Loon, R. J. D., & Prelec, D. (2022). Beta-Delta or Delta-Tau? A Reformulation of Quasi-Hyperbolic Discounting. *Management Science*, *68*(8), 6326–6335. https://doi.org/10.1287/mnsc.2022.4453

Bovens, L. (2009). The ethics of nudge. In *Preference change: Approaches from philosophy, economics and psychology* (pp. 207-219). Dordrecht: Springer Netherlands.

Clear, J. (2018). *Atomic habits: Tiny changes, remarkable results*. Avery. https://thebearchitects.com/assets/uploads/ip/atomic-habits-review-oct-2023.pdf

Cowen, T. (1991). Self-constraint versus self-liberation. *Ethics*, *101*(2), 360–373.

David, P., Kim, J.-H., Brickman, J. S., Ran, W., & Curtis, C. M. (2015). Mobile phone distraction while studying. *New Media & Society*, *17*(10), 1661–1679. https://doi.org/10.1177/1461444814531692

Duckworth, A. L. (2011). The significance of self-control. *Proceedings of the National Academy of Sciences*, *108*(7), 2639–2640. https://doi.org/10.1073/pnas.1019725108

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2016). Situational Strategies for Self-Control. *Perspectives on Psychological Science*, *11*(1), 35–55. https://doi.org/10.1177/1745691615623247

Duhigg, C. (2013). The Power of Habit: Why we do what we do and how to change. Random House.

Eyal, N. (2014). Hooked: How to build habit-forming products. Penguin.

Fogg, B. J. (2020). Tiny habits: The small changes that change everything. Houghton Mifflin Harcourt.

Frankfurt, H. G. (1988). The importance of what we care about: Philosophical essays. Cambridge University Press.

Fujita, K. (2011). On Conceptualizing Self-Control as More Than the Effortful Inhibition of Impulses. *Personality and Social Psychology Review*, *15*(4), 352–366. https://doi.org/10.1177/1088868311411165

Gabaix, X., & Laibson, D. (2017). *Myopia and Discounting* (Working Paper 23254). National Bureau of Economic Research. https://doi.org/10.3386/w23254

Gillebaart, M., & de Ridder, D. (2017). Positioning self-control in a dual-systems framework. In *Routledge international handbook of self-control in health and well-being* (pp. 35–46). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315648576-

4/positioning-self-control-dual-systems-framework-marleen-gillebaart-denise-de-ridder

Hoch, S. J., & Loewenstein, G. F. (1991). Time-inconsistent preferences and consumer self-control. *Journal of Consumer Research*, *17*(4), 492–507.

Hofmann, W., Friese, M., & Strack, F. (2009). Impulse and Self-Control From a Dual-Systems Perspective. *Perspectives on Psychological Science*, *4*(2), 162–176. https://doi.org/10.1111/j.1745-6924.2009.01116.x

Hofmann, W., & Kotabe, H. (2012). A General Model of Preventive and Interventive Self-Control. *Social and Personality Psychology Compass*, *6*(10), 707–722. https://doi.org/10.1111/j.1751-9004.2012.00461.x

Hofmann, W., Reinecke, L., & Meier, A. (2016). Of Sweet Temptations and Bitter Aftertaste: Self-Control as a Moderator of the Effects of Media Use on Well-Being. In *The Routledge Handbook of Media Use and Well-Being*. Routledge.

Infante, G., Lecouteux, G., & Sugden, R. (2016). Preference purification and the inner rational agent: A critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, *23*(1), 1–25. https://doi.org/10.1080/1350178X.2015.1070527

Jolls, C., Sunstein, C. R., & Thaler, R. H. (1998). A behavioral approach to law and economics. *Stanford Law Review, 50*(5), 1471-1550.

Laibson, D., & Maxted, P. (2023). *The Beta-Delta-DELTA Sweet Spot*. National Bureau of Economic Research. https://www.nber.org/papers/w30822

Levy, N. (2011). Resisting Weakness of the Will. *Philosophy and Phenomenological Research*, *82*(1), 134–155. https://doi.org/10.1111/j.1933-1592.2010.00424.x

McIntosh, D. (1969). *The foundations of human society*. University of Chicago Press.

Myrseth, K. O. R., & Fishbach, A. (2009). Self-Control: A Function of Knowing When and How to Exercise Restraint. *Current Directions in Psychological Science*, *18*(4), 247–252. https://doi.org/10.1111/j.1467-8721.2009.01645.x

Reijula, S., & Hertwig, R. (2022). Self-nudging and the citizen choice architect. *Behavioural Public Policy*, *6*(1), 119–149. https://doi.org/10.1017/bpp.2020.5

Rosen, L. D., Carrier, L. M., & Cheever, N. A. (2013). Facebook and texting made me do it: Media-induced task-switching while studying. *Computers in Human Behaviour*, *29*(3), 948–958.

Saghai, Y. (2013). Salvaging the concept of nudge. *Journal of Medical Ethics*, *39*(8), 487–493. https://doi.org/10.1136/medethics-2012-100727

Schelling, T. C. (1978). Egonomics, or the Art of Self-Management. *The American Economic Review*, *68*(2), 290–294.

Schelling, T. C. (1984). Self-Command in Practice, in Policy, and in a Theory of Rational Choice. *The American Economic Review*, *74*(2), 1–11.

Schelling, T. C. (1985). Enforcing rules on oneself. *The Journal of Law, Economics, and Organization*, *1*(2), 357–374.

Schuster, A. M., Cotten, S. R., & Meshi, D. (2023). Established Adults, Who Self-Identify as Smartphone and/or Social Media Overusers, Struggle to Balance Smartphone Use for Personal and Work Purposes. *Journal of Adult Development*, *30*(1), 78–89. https://doi.org/10.1007/s10804-022-09426-3

Slingerland, E. (2014). Trying not to try: The ancient art of effortlessness and the surprising power of spontaneity. Canongate Books.

Stanovich, K. E. (2004). The robot's rebellion: Finding meaning in the age of Darwin. University of Chicago Press.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The Rationality Quotient: Toward a Test of Rational Thinking*. MIT Press.

Statista. (2025, February 9). *Topic: Android – statistics & facts*. https://www.statista.com/topics/876/android/

Thaler, R. H. (2015). Misbehaving: The making of behavioural economics. Penguin Books.

Thaler, R. H., & Shefrin, H. M. (1981). An Economic Theory of Self-Control. *Journal of Political Economy*, *89*(2), 392–406. https://doi.org/10.1086/260971

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Penguin.

Tiberius, V. (2010). The reflective life: Living wisely with our limits. OUP Oxford.

Chapter 8 - Conclusion

8.1 Summary and research objectives

8.1.1 Defining digital nudging

Any investigation into nudging must navigate the contentious nature of the term itself. A broad strand of research emerged based on the two premises with which the originators of nudge, Thaler and Sunstein (2008), introduced it. That is first, a way to alter behaviour of agents by designing their choice architecture rather than by changing their incentives or choice-set. Second, Thaler and Sunstein insisted that nudges must have the aim of improving the life of the nudgee "as judged by themselves" (2008, p. 5).

The first premise states nudges are a technique to influence the behaviour of others. In addition, it describes how such influence can be exercised. The insights of how to influence originate from dual process theory and behavioural economics.

Dual process theory describes the psychology of agents as two cognitive systems, a fast and intuitive one and a slow and deliberate one. The fast and intuitive system excels at quickly making decisions on small amounts of information. Its largest drawback is that it does not operate by full comparison of choice alternatives. Rather, it operates by heuristics, which are systematically biased. As a result, the presentation or framing of available choice alternatives has a significant and predictable impact on the choice behaviour of agents. For example, most agents react differently when a shopping item is priced at €100 from when it is priced at €200 with a 50% discount.

To connect the concept of nudge more clearly with its conceptual basis of dual process theory and cognitive psychology, we adhere to a slightly modified version of Saghai's (2013) definition of nudging as A nudges B when A makes it more likely that B will ϕ, primarily by triggering B's shallow cognitive processes, while A's influence preserves B's choice-set. This definition is based solely on the

first of the two premises. We refer to it as the technical definition of nudging because it is strictly defined in terms of a technique for behaviour modification.

Behavioural economics is the study of how psychological factors influence the decisions of agents, and how these decisions systematically deviate from those predicted by mainstream, neoclassical economic theory. While dual process theory explains the psychological mechanisms, behavioural economics documents the empirical patterns these mechanisms produce. Neoclassical economics predicts that agents value a shopping item equally whether it is priced at €100 or at €200 with a 50% discount. Behavioural economists describe deviations from neoclassical predictions as behavioural "inconsistencies" (Sugden, 2018, p. 59). Most in the nudge literature have a negative appraisal of behavioural inconsistencies (Stanovich, 2011). They are often described as "errors" (Sugden, 2018, p. 14) or "inferior choices" (Sunstein & Thaler, 2003, p. 175).

If one accepts that agents indeed systematically make errors and inferior choices due to these cognitive biases, the argument for corrective intervention becomes compelling. This recognition of systematic decision-making flaws provides the normative foundation upon which the practice of nudging is justified. Consequently, many authors incorporate the second premise into their definition of nudge, arguing nudges must have the aim of improving the life of the nudgee (as judged by themselves). This phrase must be interpreted as advocating a form of ideal preference utilitarianism, as it claims agents are best off if the preference that they would have under ideal circumstances are satisfied. We refer to these nudges as libertarian paternalist nudges, in line with the terminology Thaler and Sunstein proposed.

The connection between cognitive error and corrective intervention has found resonance in policy circles and with governmental officials. Libertarian paternalists argue that government institutions are well-positioned to implement welfare-enhancing nudges, a perspective that gained significant

traction when Thaler and Sunstein became advisors to US President Barack Obama and UK Prime Minister David Cameron respectively. The persuasiveness of their argument led numerous governments worldwide to establish behavioural insight centres and "nudge units" (Halpern, 2015) dedicated to researching and implementing nudges as policy instruments.

In this dissertation, we have argued that the technical definition of nudging should be preferred over the libertarian paternalist conception for several compelling reasons. First, we have shown that there is a methodological advantage in separating what something is from its ethical evaluation. Defining nudges without built-in normative assumptions allows for a more rigorous ethical investigation. Second, it would seem odd to insist that two agents doing the exact same thing with different intentions should be labelled differently. A commercial entity employing the same psychological techniques as a government is still engaging in nudging, regardless of whether its aim is profit-maximization rather than welfare enhancement.

This technical definition enables us to examine an especially important subset of nudging practices: digital nudges. Digital nudges, as this thesis has established, represent a distinct and particularly significant category with several special features: they are delivered through digital interfaces which enables precise measurement of their effects; they can be rapidly adapted based on performance data; they can be personalized to target individual user characteristics; and they often operate invisibly, without users' awareness. These key attributes collectively define what constitutes digital nudging as examined throughout this work.

8.1.2 Identifying the ethical issues associated with digital nudging

Digital nudges are not merely incidental but fundamental to the digital economy's operation and business model. We term this phenomenon the digital nudging imperative. The digital economy functions by collecting user behavioural data and then leveraging that data to influence future behaviour. This influence is exercised most prominently in the domains of operating systems, online advertising and social media platforms, and e-commerce sites. The prevalence of digital nudges stems directly from the economic necessities of digital platforms, which depend on capturing and monetizing user attention. In an environment where user engagement directly translates to revenue, companies face strong competitive pressures to deploy increasingly effective nudging techniques.

Through our literature review, we identified several ethical issues associated with digital nudging, including concerns related to autonomy, transparency, manipulation, fairness, and welfare. After careful consideration of these various ethical dimensions, we determined that autonomy and welfare effects in the context of preference shaping represent the most pressing and foundational ethical issue requiring in-depth analysis. This focus is particularly important given that even proponents of nudging justify their interventions on welfare grounds.

Welfare concerns are especially significant because digital environments create an attention economy where users experience a constant stream of behavioural influences designed to maximize time spent, data shared, and purchases made. These influences shape not just isolated decisions but long-term behavioural patterns. These systematic patterns of influence make it increasingly difficult for users to disengage and represent a fundamental

restructuring of decision environments rather than merely isolated interventions.

Furthermore, traditional welfare analyses based on libertarian paternalism rest on the problematic assumption that while agents make errors, they can be nudged toward better choices they would make if they had fully informed preferences. We have argued that this form of ideal or purified preference utilitarianism faces significant conceptual challenges. Nevertheless, we maintain that behavioural welfare economics, despite its flawed foundations, provides valuable analytical tools for examining the welfare effects of private digital nudges when properly reconceptualized.

8.1.3 Analysing a selection of the ethical issues associated with digital nudging

To address the welfare implications of digital nudges, we suggest a model based on preference shaping, experiments in living, and reflection. In developing this framework, we have critiqued behavioural welfare economics while embracing some crucial elements from their analysis. We agree with behavioural economists that agents cannot adhere to neoclassical axioms of rational choice behaviour, which we referred to as consistency rationality. We also agree with the conceptualisation of agents as facing three fundamental limitations of imperfect information, bounded cognitive abilities, and limited willpower, which prevent them from ever attaining consistency rationality.

However, we disagree with behavioural welfare economics' solution of preference purification. This approach assumes there exists a set of 'true preferences' that can be extracted if biases are removed. We argued instead that agents often have incomplete and inconsistent preferences, raising serious questions regarding the purification project.

This alternative framework shifts from outcome rationality to process rationality. Rather than focusing on the satisfaction of hypothetically purified preferences, we propose that welfare emerges from two interdependent processes: preference shaping and preference satisfaction. Agents must first shape and continuously update their preference hierarchy before they can meaningfully satisfy preferences, and both processes must be performed proficiently. This process-oriented account embeds autonomy within welfare by evaluating not merely what agents choose, but whether they retain the capacity to reflect upon, revise, and endorse their preferences over time.

We identified two critical components of proficient preference shaping: conducting experiments in living (testing different lifestyles through direct experience) and reflecting (developing meta-preferences that guide preference formation). These components allow agents to navigate the inevitable inconsistencies and incompleteness in their preferences, turning these flaws into opportunities for learning and development.

The three fundamental limitations directly impact these processes of preference shaping and satisfaction. Improved information enables better understanding of available options and feedback from choices. Enhanced cognitive abilities facilitate reflection on experiences and preferences. Strengthened willpower supports commitment to experimentation and reflection on preferences despite immediate temptations.

This framework has significant implications for understanding how digital environments impact welfare. In profit-driven environments, digital nudges aggravate all three fundamental limitations in ways that undermine welfare. They manipulate information environments, causing information overload while also biasing information through filtering and personalization, distorting the inputs needed for preference shaping. They undermine cognitive abilities by promoting overuse of system 1, causing an overreliance on shallow processing and a scarcity of attention and reflective capacities. Additionally, they frequently

target willpower limitations through design features that encourage impulsive behaviour and immediate gratification. By recognizing how digital nudges systematically undermine the processes of preference shaping and satisfaction, we are better positioned to understand how digital environments may compromise the conditions necessary for preference shaping.

8.1.4 Recommending a program to address the selected ethical issues associated with digital nudges

To address the adverse effects of digital nudges on welfare, we argued that a user-centric approach with emphasis on education offers a comparatively promising strategy given institutional constraints for addressing the adverse effects of digital nudges on preference shaping and welfare. Both big tech corporations and governments face significant constraints in addressing the harmful effects of digital nudging. Big tech corporations are unlikely to act as attention maximisation and digital nudging are central to their business models. Many governments face jurisdictional challenges as the largest suppliers of digital nudges are all based in the US. In addition, these two stakeholder groups have already received ample attention in the literature. In contrast, users may possess both the motivation and the potential capacity to mitigate harmful effects.

In our view, users can learn to navigate digital nudge environments successfully when equipped with appropriate knowledge and tools. A focus on education is particularly promising as children and teenagers are growing up immersed in digital environments saturated with behavioural influences. Young people represent especially heavy digital users—US teenagers spend approximately seven hours daily with screens, excluding schoolwork (Robb, 2019). By providing these young users with the necessary knowledge and cognitive tools, we can enhance their digital resilience and strengthen their

capacity to proficiently shape their preferences and resist the influence of digital nudges. This approach does not eliminate the structural incentives that drive digital nudging, but seeks to strengthen agents' resilience within those environments.

Educational institutions are uniquely positioned to integrate media literacy, critical thinking, and digital self-control strategies into their curricula. Boosts (interventions that enhance decision-making competence rather than steering behaviour in specific directions) represent an especially promising tool in this context (cf. Grüne-Yanoff & Hertwig, 2016; Kozyreva et al., 2019). By equipping students with these skills early in their development, we can help them navigate digital landscapes with greater awareness and agency. Boosts align with our framework insofar as they focus on processes and aim to enhance agents' decision-making capacities rather than steer them toward predetermined outcomes.

Following this user-centric educational approach, we outline a comprehensive set of preliminary recommendations that could guide schools and educational policymakers in developing programs to equip students with skills to overcome the limitations exploited by digital nudges. These theoretical educational frameworks are organized around the three fundamental limitations and propose directions for research into strategies that might be integrated into curricula to help young people develop greater resilience in digital environments.

Information

- Develop curricula on understanding digital information environment dynamics
- Teach mindful information consumption techniques to prevent overload
- Guide students in seeking diverse and deep information sources
- Train students to customize digital tools and manage notifications

Cognitive abilities

- Integrate theoretical knowledge about digital nudges and cognitive abilities into curricula
- Develop programs focused on strengthening cognitive abilities and reflection
- Demonstrate techniques for introducing friction into digital interactions
- Practice critical thinking tools in digital contexts

Willpower

- Teach students to recognize the internal conflict between their two selves
- Train students in applying situational control strategies
- Explain the limitations of suppression as a self-control method
- Teach students to resolve temptations through pattern recognition and habit formation

8.2 Limitations and suggestions for future research

This dissertation, while offering a comprehensive framework for understanding digital nudging and its ethical implications, has several important limitations that should be acknowledged.

First, the research is predominantly western-centric in its analysis of digital nudging practices and their impacts. The dissertation primarily examines digital platforms and technologies developed by Western tech companies (particularly those based in the United States) and considers their effects on users in Western consumer contexts. This focus may not fully capture the dynamics of digital nudging in other regions, where different technological ecosystems, regulatory environments, and user behaviours exist. For instance, the digital landscape in East Asia features distinctive platforms with unique

nudging techniques that may warrant separate ethical analysis (Creemers, 2018). Similarly, the ways in which digital nudges affect users in developing economies, where digital literacy and access patterns differ significantly from Western contexts (Bovens, 2010), remain underexplored in this work.

Future research should address this limitation by conducting comparative studies of digital nudging across different cultural and economic contexts. Investigations into non-Western digital platforms such as the Chinese apps WeChat, Weibo, and Douyin and their use of nudging techniques would be particularly valuable. Studies examining the impact of digital nudges in developing economies, where digital literacy is rapidly evolving and access patterns differ from Western contexts, would also substantially enrich our understanding of the global ethical implications of digital nudging. A good starting point for future research is Selinger and Whyte's article (2011) which raises the issue of cultural variance in the effects of nudges.

Second, although the moral framework based on process rationality and preference shaping offers a promising alternative to behavioural welfare economics, its development remains relatively brief in this dissertation. The philosophical underpinnings of preference shaping deserve more thorough development, particularly regarding how different conceptions of the good life might influence our understanding of what constitutes "proficient" preference shaping. In this context, a promising next step would be a more extensive interaction with the work of Malte Dold and his collaborators, who have been developing similar approaches within behavioural welfare economics. Dold's work with Fabian (Fabian & Dold, 2022; Dold et al., 2022) on agentic preferences and his collaboration with Schubert (Dold & Schubert, 2018) toward a behavioural foundation of normative economics offer valuable insights that could further enrich our analysis. Particularly relevant is Fabian and Dold's work (see Fabian & Dold, 2022) on preference endogeneity, which directly addresses many of the concerns raised in this dissertation about preference shaping. A

more thorough engagement with this emerging literature would have strengthened the theoretical foundations of our framework and positioned it more clearly within ongoing scholarly debates concerning behavioural welfare economics.

Future research in behavioural welfare economics would benefit significantly from deeper development of endogenous preferences, process rationality, and preference shaping concepts. Grounding this moral framework in robust psychological foundations would align with best practices in behavioural economics while advancing both normative theory and practical policy applications. This approach would enable researchers to develop a more sophisticated normative framework than currently used in behavioural welfare economics and help better capture the complexities of human decision-making in digital environments.

Another potential avenue for future research might be to complement the notion of proficient preference shaping with insights from virtue ethics and Chinese philosophy. Some authors have begun exploring this direction. Matson and Dold (2021) for example discuss David Hume's ideas on preference shaping, and his indebtedness to stoic philosophy. A similar argument could be made for John Stuart Mill. In addition, Edward Slingerland (2014) connects dual process theory to ancient Chinese philosophy and his analysis opens many interesting lines of thought regarding how preferences could be shaped proficiently. Particularly, his discussion shows the relevance of ancient Chinese thinkers for the discussions on willpower, habits, and cognitive abilities and the interaction between system 1 and 2.

Third, this research focuseson philosophical roots, rather than the neurological underpinnings of decision-making and nudging. Although we incorporated insights from the growing field that neuroeconomics offers, it offers potentially many more valuable insights into how digital nudges affect neural processes, which could provide a more robust biological foundation for

understanding their impacts. Incorporating these neurological perspectives might strengthen both the descriptive account of digital nudging and the normative analysis of its implications. This limitation is especially relevant since many of the psychological concepts in this work, for example system 1 and 2, cognitive abilities, willpower, attention, would be strengthened by strong neurological underpinnings.

Fourth, the conceptual status of system 1 and system 2 remains somewhat contested. While these concepts serve as useful analytical tools, questions about whether they represent actual cognitive systems, "sets of systems" (Stanovich, 2011, p. 28), or metaphorical constructs is not fully resolved. This point becomes especially relevant in chapter 7, where we align with the literature and speak of two selves. Greater conceptual clarity regarding these foundational psychological constructs would strengthen the theoretical framework of the dissertation and the research field of behavioural economics in general.

Fifth and finally, while a user-centric educational approach is proposed as the most promising strategy for addressing the ethical issues associated with digital nudging, the pedagogical and psychological foundations of this approach require further development. More detailed consideration of educational theories, teaching methodologies, and evidence-based interventions would enhance the practical utility of the recommendations. Additionally, the effectiveness of the proposed educational interventions should be empirically evaluated to ensure they achieve their intended outcomes.

Future research should develop and empirically test specific educational interventions designed to mitigate the harmful effects of digital nudging. Longitudinal studies measuring the impact of digital literacy education on students' ability to navigate digital environments would be particularly valuable. Experimental studies comparing different pedagogical approaches for teaching about digital nudging could help identify the most effective educational strategies. Additionally, research examining how educational interventions

might be tailored to different age groups and cultural contexts would increase the practical applicability of the recommendations. Collaborations between philosophers, psychologists, and education specialists could yield more integrated approaches that address both the ethical principles and practical implementation challenges of digital nudging education.

In conclusion, this dissertation has established that digital nudges significantly harm human welfare by undermining proficient preference shaping across the three fundamental limitations of information, cognitive abilities, and willpower. Our process rationality framework offers a more robust alternative to traditional behavioural welfare economics by recognizing the importance of preference shaping alongside preference satisfaction. The educational approach we propose acknowledges that we should not expect meaningful changes from big tech companies, as addressing these issues would undermine their core business models. Instead, focusing on students and equipping them with skills to enhance their informational awareness, cognitive abilities, and willpower will advance welfare as people learn to navigate digital nudges.

8.3 Bibliography

Bovens, L. (2010). Nudges and Cultural Variance: A Note on Selinger and Whyte. *Knowledge, Technology & Policy*, 23(3–4), 483–486. https://doi.org/10.1007/s12130-010-9128-2

Creemers, R. (2018). China's Social Credit System: an evolving practice of control. Available at SSRN 3175792.

Dold, M. F., & Schubert, C. (2018). Toward A Behavioral Foundation of Normative Economics. *Review of Behavioral Economics*, *5*(3–4), 221–241. https://doi.org/10.1561/105.00000097

Dold, M., van Emmerick ,Elias, & and Fabian, M. (n.d.). Taking psychology seriously: A self-determination theory perspective on Robert Sugden's opportunity criterion. Journal of Economic Methodology, 1–18. https://doi.org/10.1080/1350178X.2024.2402981

Fabian, M., & Dold, M. (2022). Agentic preferences: A foundation for nudging when preferences are endogenous. *Behavioural Public Policy*, 1–21. https://doi.org/10.1017/bpp.2022.17

Grüne-Yanoff, T., & Hertwig, R. (2016). Nudge Versus Boost: How Coherent are Policy and Theory? *Minds and Machines*, *26*(1–2), 149–183. https://doi.org/10.1007/s11023-015-9367-9

Halpern, D. (2015). Inside the nudge unit: How small changes can make a big difference. Random House.

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, 21(3), 103-156.

Matson, E. W., & Dold, M. (2021). The Behavioral Welfare Economist in Society: Considerations from David Hume. *Review of Behavioral Economics*, 8(3–4), 239–258. https://doi.org/10.1561/105.00000141

Robb, M. (2019). *Tweens, teens, and phones: What our 2019 research reveals*. Common Sense. https://www.commonsensemedia.org/kids-action/articles/tweens-teens-and-phones-what-our-2019-research-reveals

Saghai, Y. (2013). Salvaging the concept of nudge. *Journal of Medical Ethics*, *39*(8), 487–493. https://doi.org/10.1136/medethics-2012-100727

Selinger, E., & Whyte, K. (2011). Is There a Right Way to Nudge? The Practice and Ethics of Choice Architecture. *Sociology Compass*, *5*(10), 923–935. Scopus. https://doi.org/10.1111/j.1751-9020.2011.00413.x

Slingerland, E. (2014). Trying not to try: The ancient art of effortlessness and the surprising power of spontaneity. Canongate Books.

Sugden, R. (2018). The community of advantage: A behavioural economist's defence of the market. Oxford University Press.

Stanovich, K. E. (2011). *Rationality and the reflective mind*. Oxford University Press.

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *American economic review*, *93*(2), 175-179.

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Penguin.