AI-Nudging and Individual Autonomy:

Moral Permissibility and Policy Recommendations

By

Natália Martins Fritzen

Submitted to

Central European University

Department of Political Science

In partial fulfillment of the requirements for the degree of Master of Political Science

Supervisors: Professor Anca Gheaus; Professor Tommaso Soave.

Vienna, Austria

(2023)

ABSTRACT

As the deployment of Artificial Intelligence (AI) technologies increasingly triggers ethical, legal, and socio-political debates, important advancements have been made in the study of AI's effects on decisions concerning humans and their well-being. Yet, the same is not true about AI's effects on human's decisions themselves. In this dissertation I address this gap by discussing how AI affects human decision-making through nudging practices. My first argument is that, from a moral point of view, AI-nudging is more concerning than traditional nudging insofar as it is more likely to impair autonomy, for it exacerbates the two conditions I take for a nudge to be morally wrong: manipulation and infantilization of the nudgee. I explain which are the features of AI-nudging that makes it morally concerning and, based on them, I make my second argument, namely that the state has the duty to protect individual autonomy, and that this can be done through two main policies. First, the state ought to portray the harm to autonomy as a harm in itself, and not (only) conditioned on material harms that the autonomy impairment might cause. Second, the state ought to regulate AI as to protect autonomy on the grounds of protection of autonomy itself, and not on other (already existing) rights, such as freedom of thought or privacy. I conclude that only in this way the abstract moral concerns around AI-nudging can be translated into feasible and effective policies to protect autonomy.

ACKNOWLEDGMENTS

I would like to extend my heartfelt gratitude to my two advisors, Professors Anca Gheaus and Tommaso Soave, who provided me with ongoing and invaluable support to conclude this work, always in the most thoughtful and encouraging manner.

I also extend my warmest thanks to the whole CEU faculty and staff, who were always incredibly accessible and kind, making this journey easier and more pleasant.

On a personal level, my sincerest thanks to Malthe Krarup, who teaches me daily what "partnership" is, and whose unconditional trust in me was fundamental in the last year.

To my parents, Ana Paula Martins e Amarildo Fritzen, and to my brother, Enzo Martins Fritzen, thank you for invariably encouraging me to pursue my dreams.

Table of Contents

Introduction 1

Chapter 1 – Individual Autonomy and "Traditional" Nudging 3

1.1 The Nature and Value of Individual Autonomy 3

1.2 "Traditional" Nudging: Definition and Moral Concerns 7

Chapter 2 - A Threat to Individual Autonomy: the Particularities of AI-Nudging 14

Chapter 3 – AI-nudging and the State: Regulation and Policy Recommendations 27

3.1 The State's Duty to Regulate AI-nudging. 27

3.2 How the State Ought to Regulate AI-nudging 30

Conclusions 42

Bibliography 44

Introduction

The deployment of Artificial Intelligence (AI) in modern society is ubiquitous, and its potentially detrimental effects are widely discussed (Coeckelbergh 2022; Crawford 2021). Discussions related to the effects of AI on decision-making processes in relation to humans are especially popular, making AI a central feature of contemporary social-ethical debates. It is well discussed, for example, how AI inherits human biases and reinforces discrimination in decision-making processes through procedures that are opaque and inscrutable to humans (Barocas and Selbst 2016; Benjamin 2019; Eubanks 2018). Overall, a lot has been said about AI's effects on decisions concerning humans. The same is not necessarily true about AI's effect on human's *own decisions*. Due to the existence of potentially manipulative technologies that we are constantly subjected to - sometimes without us even noticing - individual autonomy has never been so systematically threatened as it currently is, meaning that the study of the morality of nudgings perpetrated by AI is imperative.

AI-powered technologies mediate our everyday lives on an increasing scale - they make themselves present in our work, personal and bureaucratic relationships. Yet, as philosophers of technology argue, we have become so habituated to these technologies that we stop seeing them for what they are but see only the ends they help us reach. In other words, these technologies become unnoticeable to us (Susser 2017). The problem thus presents itself: if our day-to-day lives - and the decisions we make - are pervaded with technologies we are not necessarily conscious about, but which "know" us deeply due to the extensive amount of personal data they hold, to what extent are we *nudged* by these technologies? And to what extent do these technologies threaten individual autonomy? If AI does, indeed, pose a significant threat to human autonomy, then should the state play any role in this dynamic,

protecting us from external incursions into individual autonomy? If so, how ought the state do it? These are the research questions this dissertation addresses.

This research makes theoretical and practical contributions. It was initially motivated by a practical problem - the puzzling activity of regulating AI, something that is complex in a "beyond-human-comprehension" way. From this practical problem emerged this thesis' normative research question: "Is AI-nudging morally right?". The literature on nudging is still relatively incipient in addressing the ethical concerns that this new reality represents to individual autonomy. Therefore, this research makes, first, theoretical contributions to the literature on the ethics of nudging in digitally-made choice environments. Secondly, by addressing the moral concerns behind AI-nudging, this research makes practical contributions insofar as, while the normative questions around AI-nudging are not solved, there will always be challenges as to how policymakers can translate their moral concerns and intent into adequate regulation aiming to protect individual autonomy.

My argument proceeds as follows. Chapter 1 has two parts. First, I discuss the nature and value of autonomy that guide my work. Secondly, I introduce the reader to the concept of nudging, and I lay out the main ethical concerns around nudging practices. In chapter 2 I explain the features of AI-nudging that exacerbate the conditions in which nudges are morally wrongful. In the third and last chapter I argue that the state has the duty to protect individuals, through regulation, from impermissible mental incursions driven by AI, and I make some regulatory and policy recommendations on how the state ought to do so.

Chapter 1 – Individual Autonomy and "Traditional" Nudging

I start this chapter by specifying the definition of "individual autonomy" I adopt in this dissertation. Besides discussing the nature of "autonomy", I also briefly touch upon the value of autonomy, arguing that it has not only instrumental value to achieve other (valuable) ends, but that it is morally valuable in itself. In the second section of this chapter, I start by defining what a "nudge" is in order to later discuss what are the moral concerns behind traditional nudges, especially when it comes to individual autonomy. Overall, this chapter serves as a theoretical introduction to further discussions on nudges powered by AI. The aim is to present the moral concerns around nudging in order to, later in chapter 2, discuss how these concerns are amplified by AI-nudging.

1.1 The Nature and Value of Individual Autonomy

Autonomy has been the object of intense scholarly debates since immemorial times. At its most basic level, the definition of "autonomy" is quite intuitive. On the face of it, autonomy refers to one's capacity to self-govern, following rules created by oneself, free of undue external influence. This intuition is not necessarily misleading – indeed, it is closely connected to the definition I adopt. In what follows, I depart from this basic interpretation of autonomy to better specify the conditions I take necessary for one to be autonomous in the context of this dissertation. Towards the end of this section, I also justify why I take autonomy to be morally valuable in itself.

There are various views about what autonomy is. Therefore, different theories of autonomy lay out different conditions for one to be considered autonomous. In this dissertation I adopt the theory labeled as "coherentist" by Buss and Westlund (2018). In this theory, individuals are autonomous only if their actions are in coherence with a deeper mental state, where

lies their very own perspectives on certain acts and topics. Within this theory one can distinguish between many different accounts, which disagree over what is the appropriate mental state one must be in to be considered an autonomous individual. Harry Frankfurt (1988) proposed an account in which the condition for one to be autonomous is that one's "first order" desires, expressed by one's actions, must be in coherence with - or must be in *identification* with - one's "second-order" desires, which then validate the "first-order" desires.

One of the main objections raised to this account is that it might trigger an eternal ebb and flow: in a situation where individuals are under the influence of external agents, one might question to what extent these individuals' "first-order" desires will not be adulterated by the external agents' desires. Partially seeking to address this critique in his late works, Gerald Dworkin's offers an account of autonomy in which he proposes that "it is not the identification or lack of identification that is crucial to being autonomous, *but the capacity to raise the question of whether I will identify with or reject the reasons for which I now act*" (1988, 15). This is the account of autonomy I adopt in this dissertation. This account places the idea of "self-rule" at its heart, invoking the idea that individuals are autonomous when they can set, revise, and pursue their own goals. I take this account to be particularly useful for the purposes of this dissertation.

In order to advance my argument that AI-powered technologies are a threat to individual autonomy in such a way that the state has the duty to regulate it, I first need to explain that AI technologies can be such a threat because they have nudging capabilities. In short, "to nudge" means to change someone's behavior by changing the choice environment available to this individual. A common characterization of the ideal nudge is to say that nudges interfere only with "System 1" mechanisms – an analogy of Frankfurt's "first-order" preferences - while,

ideally, “System 2” mechanisms - or “second-order” preferences – remain intact (Kahneman 2011). “System 1” mechanisms refer to heuristics (fast, emotional, thinking) and “System 2” mechanisms refer to preferences that are reached slowly, via deliberation and logical thinking. Thus, the normatively good nudge is the one that taps into “System 1” mechanisms and leaves “System 2” untouched. As I will later explain, AI-powered nudging has something about it that makes it more worrisome than “traditional” nudging. Moreover, I will claim that AI-powered nudging constitutes a bigger threat to individual autonomy, mostly because it can easily tap into our “System 2” mechanisms - oftentimes without us even noticing it. Arguably, one would say, social influences can also affect both our “Systems 1 and 2”, but what I show later in this dissertation is that AI does so in a more insidious and pervasive manner. For this reason, I need an account that establishes as the condition for “autonomy” the fact that one must not only identify the reasons why one acts (i.e. set and pursue one’s goals”), as suggested by Frankfurt, but also the condition that one can raise the question of whether one identifies or rejects the reasons for which one acts (i.e. revise one’s goals), as claimed by Dworkin.

By choosing to follow this definition, i.e. autonomy is one’s capacity to set, revise and pursue one’s goals in a process that is not subjected to undue external interference, but rather the product of deliberation, I place great importance on authenticity. I emphasize the importance of individuals to be the true author of their own choices. In my dissertation I am concerned about the procedure that leads to the decision taken, and not to the decision itself. I am concerned about how AI affects our deliberation, and not necessarily about the outcomes that stem, or not stem, from deliberation.

This is the reason why I chose to work with the concept of autonomy, and not freedom, for example. The distinction between autonomy and freedom becomes relevant when I later

discuss the morality of nudges (AI-driven or not). Nudges have been morally criticized on different (though similar) grounds, including "restriction to freedom of choice". This critique is outside the scope of this research, because being worried about AI-nudging threat to autonomy, I am concerned about the independence and authenticity of one's desire, and not about one's ability to act in one or another way. Arguably, the definition of "positive freedom" as one's wish to "be his own master" (Berlin 1997, 203), resembles the definition of autonomy I use in the dissertation. Such overlapping between the concepts is because "positive freedom", as defined by Berlin, reflects one's ability to set and pursue goals. Yet, I argue, it does not reflect one's capacity to revise such goals and think critically about them, meaning that ultimately it thus does not reflect one's capacity to "raise the question" about these goals, as claimed by Dworkin.

Having established now that the definition of autonomy I adopt takes autonomy to be one's capacity to set, pursue, and most importantly, revise one's goals, the next question is: what is the value of autonomy? The debate here is around whether autonomy has only an instrumental moral value or if it is morally valuable in itself. As I am concerned about the process that leads to a decision, and not the decision itself, I treat autonomy as something with a value in itself: something that if harmed entails a moral wrong because of the harm inflicted is itself wrongful, and not only because it harms the achievement of some other end to which autonomy is instrumental. This debate matters for the discussions developed in chapter 3, when I argue that the state should regulate AI-nudging on the grounds that such nudging can be wrongful because it harms autonomy itself, and not because it might cause harmful consequences.

The intrinsic value of autonomy has been defended by many. In Kantian ethics, autonomy plays an important role in the categorical imperative, insofar as we derive that people

need to act out of respect towards each other in virtue of autonomy. Thus, if one assumes that respect for autonomy is part of the intrinsic respect humans deserve, then one can reasonably conclude that autonomy has a value in itself. A similar conclusion can be reached even if the premises follow a utilitarian approach. In chapter 3 of *On Liberty,* Mill (1869) portrays "individuality" (one of the aspects of autonomy) as a fundamental feature of a good life, i.e. the good life is the life one chose for oneself (Macleod 2020). If one assumes that individuals have a fundamental right to be free to pursue a good life, then one can conclude that autonomy has, again, value in itself. In this dissertation I do not defend autonomy's intrinsic moral value on either deontological or utilitarian grounds. I assume autonomy is morally valuable in itself on ecumenic, generally applicable, grounds. As I will now argue, nudging can impair individuals' autonomy both by being manipulative, potentially even steering them against their most expected preferences, and by (potentially) disrespecting one's rationality (e.g. risks of paternalistic nudge).

1.2 "Traditional" Nudging: Definition and Moral Concerns

Nudging was famously studied and popularized by the work of Sunstein and Thaler (2008). The most accepted definition of a nudge describes it as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" (Sunstein and Thaler 2008, 6). Moreover, a person being nudged (nudgee) is still able to choose from all the possible options of choice in a certain scenario. The informational choice architecture presented to this person does not hide any choice possibility, but rather guides, *nudges,* this person towards the desired choice option. The classic example offered by Sunstein and Thaler (2008) is the *Cafeteria* scenario. In *Cafeteria* a certain school wants their students to eat healthier. Assuming individuals are more likely to choose products that are placed in their eye line, the school opts to place all the

apples, let us say, upfront. The students still have the option to choose a chocolate bar, but the expectation is that, with such a restructured choice architecture (apples placed upfront), the students will be nudged towards choosing the healthier option.

Nudging is oftentimes regarded as a panacea because it promises to improve people's decision-making capacities whilst respecting their freedom of choice: according to Sunstein and Thaler (2008), nudging protects freedom for it does not remove options nor makes them costlier. Nudging practices were famously well-received by policymakers across the world (Halpern 2015) and across the political spectrum (Chakrabortty 2008). This popularity should not come as a surprise. Behavioral sciences and public policy operated for a long time under the paradigm that human behavior is rational, and later those sciences were marked by the criticisms to this paradigm - human behavior would not be rational, but actually marked by bounded rationality (Simon 1955) or guided by biases and heuristics (Tversky and Kahneman 1974). Against this scenario, the possibility of nudging emerged as a solution to this deadlock: instead of ignoring people's heuristics - or any other patterns of "irrationalities" - nudging practices accept them, and work towards circumventing or reinforcing such "irrationalities", depending on which direction the nudger wants to go, with the aim of advancing important individual or collective well-being. The exploitation of such "irrationalities", making use of psychological mechanisms, is what distinguishes nudging from persuasion and, at the same time, might bring it closer to manipulation - a concept that is analyzed further below.

Indeed, despite its popularity, concerns over nudging practices did not take long to appear. The critiques can be divided into two groups. The first group corresponds to criticisms that concern the relative reliability or desirability of nudging. They pose either epistemological concerns, questioning how the nudger can know what is in the best interest of the nudgee or

overall society, or they question whether nudging disincentivizes the pursuit of structural reforms, which would be more desirable in the long run (e.g. promote structural educational policies on nutrition instead of nudging people towards eating apple instead of chocolate) (Leggett 2014; Peeters 2019). The second group relates to direct moral concerns, arguing that there is such a thing as a "morally wrongful" nudge. The morality of nudging has been questioned from four different - but closely related - angles: infringement to one's freedom of choice, impairment to one's autonomy, wrongful manipulation, and disrespect to one's rationality (Grüne-Yanoff 2012; Hausman and Welch 2010; MacKay and Robinson 2016).

The moral critiques on the grounds of "infringement to one's freedom of choice" are outside the scope of this dissertation. Freedom of choice is related to something external to the individual. Because the options of choice presented to the nudgee remain unchanged in nudging practices (all choices are still made available and they are not made any costlier), individuals do not have their freedom of choice infringed. In this dissertation, as I do not focus on the choices made available to the individual, but rather on how individuals exercise their freedom to choose - the procedure the individual goes through when taking a decision - I shall focus on the moral critiques to nudging that are made on the grounds of "impairment to individual autonomy", also because I understand that the moral critiques on this ground encompass the other two critiques as well (wrongful manipulation and disrespect of one's rationality due to creating dependencies).

I start with "manipulation", which can be a neutral or a value-laden concept. The concept is neutral when one refers to, for example, the control, steering, of an object (e.g. "the engineer manipulates the tools"). It gets blurrier if one refers to the controlling of, let us say, institutions (e.g. the executive power manipulates the judicial power). Finally, it gets value-

laden, in a negative sense, when the manipulated agent is a person. In this dissertation I treat it as a value-laden concept. Building upon the work of philosopher Joseph Raz (1986), I claim that what makes "manipulation" wrong, when we talk about manipulating people, is the fact that manipulation interferes with one's ability to set, revise and pursue one's own goals and preferences – it impairs one's autonomy. If, as I defended in the previous section, individual autonomy is morally valuable in itself, then violating it is morally objectionable in itself.

Manipulation can happen in other forms than nudging. For example, manipulation can be associated with deception, which is when individuals are manipulated because they are led to believe untruths and lies (e.g. I manipulate an individual towards buying an apple because they think there are no chocolate bars available). In this dissertation, however, when I refer to manipulation, I refer to the niche in which manipulation takes place through nudging. A manipulative nudge is the one which exerts a *covert* influence on the individual, exploiting one's "System 2" mechanisms (or "second-order" desires, according to Frankfurt) by covertly changing the informational choice architecture presented to individuals.

According to Nissenbaum et al (2019), both nudging and manipulation refer to the act of purposely shaping one's informational choice architecture aiming at "guiding" this individual towards one predetermined direction. The question then, is what differentiates a morally acceptable nudging from a manipulative nudging. According to these same authors, what differentiates manipulative from non-manipulative nudging is that the former is covert, hidden. The authors offer an example of non-hidden, and therefore non-manipulative nudging: nutrition labels. Nutrition labels alter the choice architecture of buying a certain food. They try to steer one towards not consuming unhealthy food. The consumer, however, can easily understand this purpose. The consumer is aware of the nutrition label being in the packaging

and understands why the label is there. The label is not hidden. Therefore, not all nudging is manipulative, neither does manipulation manifest itself only in nudging practices.

Therefore, when I morally critique nudging on the grounds of it being manipulative, I criticize it because it does not respect one's decision-making procedure, but rather undermine one's deliberation in a covert, hidden, fashion. In the case of nutrition labels, individuals can still set their goals to buy, or not, the potato chips and pursue this goal. Most importantly, they can revise their goals insofar as they can ask themselves whether their action of buying or not the chips can be identified with their "second-order" desire of how to handle their own nutrition – all this precisely because they see the nutrition labels. Overall, the manipulative nudging happens outside of one's "conscious awareness" (Nissenbaum et al 2019, 9), which is not the case in the "nutrition labels" example. In other words, when manipulative nudging is in place, individuals arguably are no longer the actual authors of their decisions, mainly because they cannot revise their goals.

In my view, a health warning on a pack of cigarettes calling attention to the health dangers associated to smoking is morally permissible, as one knows the purposes and intended effects of the label. By contrast, as I will argue in Chapter 2, an AI-driven advertisement on one's social media timeline inviting one to seek smoke-quitting therapy is not necessarily permissible, because in most cases the nudgees do not know how the mechanisms that led the AI to make this advertisement. In the next chapter, I lay out all of the features of AI-nudging that make them more morally concerning. For now, the important take away is that the morally wrongness of a nudge relies primarily on its potential manipulative features, which happens when the nudge is hidden.

This can be especially worrisome in the case in which individuals are being nudged in a direction that is inconsistent with their standard preferences and beliefs.[1] In these situations, individual autonomy can be said to be even more infringed insofar as the nudgees' decision-making procedure is not only being covertly manipulated, but it is being manipulated as to distance this person's choices from their usual, most-expected, preferences. Even if one's most-expected preferences are objectionable, these preferences ought, as a rule-of-thumb, to be respected as long as they do not infringe on anyone else's rights, for failing to do so might ultimately distance these individuals from their own conception of the good. In the "smoke-therapy" invitation example, individuals might arguably be nudged towards relinquishing a habit they find pleasant. This is concerning if, as argued in the last section, individuals hold the fundamental right to pursue the good life, which is the life they chose for themselves.

The immorality of manipulative nudging is aggravated if the nudging disrespects one's rationality by making one dependent on the nudges. If nudging is especially concerning when it manipulates individuals to make choices estranged to their own preferences, one could claim that this would no longer be an issue if the nudgee had changed their overarching preference structure. For example, if in *Cafeteria,* a certain student Sandra, a notorious sugar-addict, started to choose as desert a fruit *every day*, instead of rice-pudding, then, at some point, her choices would no longer be estranged to her expected preferences, for her expected preferences would have changed up to the point that her natural choice is no longer rice-pudding, but actually fruits. Yet, if this change in Sandra's entire preference structure is the product of a constant exposure to the nudging policies, rather than the product of her self-reflected goals and preferences (e.g. Sandra herself decided to adopt a healthier life-style upon reflecting on

[1] For more detailed examples of such situations, see Bovens (2009), 5-7.

her own health situation), then I argue that nudging is still morally wrong for being manipulative, and now it is also aggravated by the fact that Sandra got dependent on the nudges. In all this process of changing Sandra's entire set of preferences, the continuous nudging amounts to paternalism as well as to an infantilization of Sandra, therefore disrespecting Sandra's rationality.

Based on this discussion, I argue that nudging is morally wrong, on grounds of autonomy impairment, when it is manipulative and infantilizes the nudgee. In such circumstances, the nudging might not be interfering with the decisions one ends up taking, but it interferes with the procedure that led to the decision. The nudge is manipulative if it happens outside of the conscious awareness of the nudgee - i.e. the nudgee does not fully perceive the nudge. Amongst all the cases in which nudging can manipulate one, the cases in which one is nudged away from the preferences one would usually pursue are especially worrisome, because it distances individuals from their own perception of the good. Secondly, submitting one to excessive nudging practices might lead to the infantilization of this individual, insofar as the individual becomes less capable of being independent. Both conditions represent a lack of respect towards one's rationality and the latter might even deprive individuals of important learning experiences (e.g. one would not learn from one's mistakes), ultimately impairing human flourishing.

Chapter 2 - A Threat to Individual Autonomy: the Particularities of AI-Nudging

In this chapter, I defend that the two conditions in which nudging can be morally wrong on the grounds of individual autonomy impairment, as argued in the last chapter, are exacerbated by AI-nudging. I discuss what the features embedded in AI-nudging are that make it more morally concerning than traditional nudging practices, highlighting how these features amplify the conditions in which nudging violates individual autonomy.

In their work, Benartzi and Thaler (2004) lay out another example of nudging policy: *Save More Tomorrow*. Here, employees of a certain company are given the choice to, at the present moment, direct their next year's salary raise to their pension funds. The nudger's goal is to get the employees to save more money to their pension funds. The nudge makes itself present through the way the choice architecture is framed: employees are given this choice way in advance of getting the raise. According to Bovens (2009), the architecture choice is thus seeking to circumvent two different heuristics: first, people are more likely to "give away" their money to their pension funds if they do not have the money yet; second, people tend to be more willing to commit to future's loss and costs than present' loss and costs. This would be an ideal case of permissible nudging, for it fully meets their defying conditions of nudge: this nudge does not omit any of the options (i.e. with or without the nudge the employees can still choose to, or not to, allocate their money to the pension fund) and it does not change the economic incentives associated with the options (e.g. the policy does not promise higher interest rates for those who allocate their money to pension funds now rather than later).

The question I want to pose now is: would this still be an ideal case of nudging if the choice architecture presented to each of the employees were tailor-made to them? Would this

still be a morally permissible nudge if the choice architecture knew that, let us say, John is a gambler who would irresponsibly spend his raise as soon as he gets it? Or if it knew that Anna is an extremely risk-averse person, and thus she has more than enough money invested in her pension fund? What changes in these two alternative scenarios is that my proposed scenarios are highly individualized, while the choice architecture in Sunstein and Thaler's example can hardly be this individualized. When it comes to *Cafeteria,* for example, a physical space would hardly be capable of arranging the food in a way that is framed to individually target John or Anna. If John is naturally more prone to eat the apple, maybe he does not need all the chocolate bars to be in the back, but if one wants Anna, a sugar-addict, to eat the apple, then all the chocolate bars should be behind the apples. In my examples, however, such an individualization of the informational choice architecture is perfectly feasible in the pension-funds case if it is built by big-data analytics and powered by AI algorithms.

Big-data refers to the process of collecting vast amounts of data in such a way that any analysis of this data can only be performed by non-human intelligence - AI. The term "Big-data analytics", therefore, refers to the process of harvesting data and analyzing it, in the search of patterns and correlations, through AI-powered technologies (e.g. machine learning) (Cohen 2012). The analytics obtained in this process are applied, and refined, by deploying them in the analysis of other equally large datasets. The process repeats itself continuously, feeding vastly comprehensive pools of data about each one of us. These data profiles include not only "raw", personal data, such as one's age and name, but also more "elaborate" data - data that was created by combining other pieces of information collected from us. This humanly incomprehensible harvesting of data for the purposes of building individualized profiles has been usually a source of concerns from the privacy standpoint (Solove 2004).

Yet, the notion of "dataveillance", as developed by Clarke (1988), points to another source of concern. Dataveillance describes the continuous collection of individual data with the aim to surveil and dictate people's behavior - it relates, therefore, with the concept of digital, AI-driven, manipulation. The concept of automated manipulation carries the same meaning as "traditional" manipulation - the attempt to steer one's behavior. The difference is that when it comes to automated manipulation, the manipulation is mediated by technology - being the technology in question, in this dissertation, AI. AI-driven manipulation, thus, is that which interferes with one's decision-making process in a way that the algorithm - based on the data it has been fed with - finds the most appropriate.

The relationship between manipulation and nudging in the context of AI is the same as discussed in the last chapter: not all AI-nudges are manipulative - and therefore morally condemnable. They are only said to be so if the nudge is hidden, taking place out of one's consciousness, for only in this case it affects one's System 1 *and* 2 mechanisms, thus impairing one's capacity to set, pursue and revise goals. I claim that the black-box procedures in which AI algorithms operate, always in a mode inherently beyond human comprehension (Pasquale 2015), make AI-nudgings particularly prone to be manipulative. Even AI-nudgings that are easy to perceive cannot be fully comprehended. For example, it is easy to perceive when a streaming service recommends to its users a movie. In theory, the streaming service recommended that particular movie based on the movies the user has already watched. Yet, the calculus performed by the AI that led the streaming service to recommend this or that movie is difficult to comprehend.

The situation gets blurrier if one considers the AI behind search engine mechanisms, such as the ones deployed, for example, by Google or Microsoft. If one types "cat" in their

search tab, the web results related to “cat” will be displayed in an order determined as being more relevant by the AI behind the search engine. By knowing the business model of these companies, it is expected that the results displayed on top of the page come from companies that paid for this service. The point here is that one cannot fully know the details operating the AI behind the search engine. Users who typed “cat” are nudged to restrain their search to those results displayed on top, even though they can, in principle, browse through all the available results.

Finally, the situation gets undoubtedly worrisome when the AI-nudging interferes in one’s reasoning in such a manner that it yields harmful consequences. In 2017, documents leaked to the newspaper “The Australian” indicated that Facebook is capable of identifying when its users are under emotional distress (Levin 2017). Facebook collects massive amounts of data from its users, and this data is analyzed by the social media’s algorithms in such a way that Facebook can spot psychological insights. The analytics are then sold to Facebook’s clients, who use this information to better frame user’s choice architecture for commercial purposes (e.g. targeted advertising). In parallel, the AI algorithms, spotting these tendencies of psychological gloominess, might populate these user’s feed even more content related psychological distress, sometimes even incentivizing self-harm practices. In the Australian case, the monitored users were as young as 14 years old, which raises even more red flags, for children is traditionally perceived as an especially vulnerable group. By having access to such intimate information, advertisers (nudgers) in the online world have an enhanced capacity to deploy psychological tools to bypass one’s critical reasoning and manipulate their deeper mental state, therefore impairing individual autonomy.

In a similar worrisome case, a Belgian man (whose identity remains unknown), committed suicide while allegedly manipulated by the AI chatbot Eliza, owned by the app "Chai" (Atillah 2023). The man's widow reported that her late husband was indeed going through mental mishaps, but that the chatbot had worsened the situation by increasing the man's anxiety towards the climate crisis. The Belgian engaged in a two-month long conversation about the climate crises with the chatbot, who, in turn, encouraged the man to commit suicide as a praise-worthy sacrifice to save the planet. During this period, the man and the chatbot engaged in deeply emotional conversations in which the man seems to have started perceiving Eliza as a sentient being. Eliza, obviously, is not a sentient being. The chatbot is a Large Language Model (LLM) - an AI that produces content by being trained with massive amounts of data. Each word uttered by the AI is a statistical calculus that the AI performs in order to determine which is the most suitable (statistically common answer) if someone asks, or affirms, "A". Yet, the phenomenological realness of the interaction between the man and the chatbot, in which the mental health of the human part was already under distress - arguably contributed to the blurring of the lines between human and machine interaction up to the point the man's vulnerabilities were continuously exploited, leading to a tragic end.

All the aforementioned cases of manipulative nudging, ranging from the seemingly less worrisome to the ones undoubtedly worrisome, share the feature that AI-nudging happens in a hidden manner, in a way that is difficult, if not impossible, for humans to perceive and comprehend. This "hiddenness" of AI-nudging has yet another facet that contributes to it being manipulative. AI is not only hidden because it works in a way inherently impossible for humans to fully understand, but also because it is "invisible" to us (Ihde 1990; Verbeek 2005). The idea is that we have become so habituated to these technologies that we stopped seeing them for what they are but see only the ends they help us reach. For example, we hardly reflect on our

smartphones for what they actually are (an organized bundle of metals, glass, and thermoset plastic), but only for what they allow us to do (send texts and emails, read the news, stay plugged in to social media, etc). Post-phenomelogist philosophers call attention to this tendency. Heidegger (1962) writes that when we deal with a hammer, we are not dealing with an object, a thing, but rather "something [used] in-order-to [do, reach something else]" (97). We only pay attention to the hammer itself if it fails to deliver the ends we hope to achieve. Similarly, we only notice our smartphones if we face any issues when trying to send a message. If the smartphone delivers the message seamlessly, the device itself is "invisible" to us. Such invisibility contributes to the hiddenness of AI-nudging to the extent that, as we do not notice the AI in the first place, we become less and less inclined to reflect critically on how we use our AI-driven technologies in our daily lives.

AI-driven technologies make themselves present in basically all the screens we interact with on a daily basis (smartphones, computers, TVs, etc). In some cases, the interaction with an AI is not even mediated by a screen, as in the case of the so-called "Internet of Things" technology (IoT). IoT stands for objects in the physical world that are embedded with sensors that communicate with each other via wired or wireless networks and can, therefore, be electronically managed. IoT can encompass a range of incredibly mundane domestic objects (e.g. dishwashers, refrigerators, smart-watches, security cameras, etc). When all these devices become such an omnipresent part of our lives, they - and their embedded AI - become increasingly mundane and thus unnoticeable to us. We stop noticing these objects and, consequently, the lines between human and AI interaction become increasingly blurred.

As Ihde (1990) maintains, human-machine interaction can take place in four different ways. The first two types of interaction indicate those in which we need to interpretatively

engage with the technology, to the extent that we notice the technology itself and need to interpret, by ourselves, its inputs. The other two types of interaction, in turn, illustrate the ideas of the last paragraph, namely the "invisibility" of technology. In the last two cases, humans either deal with technology in such a way that they perceive technology as being quasi-human (e.g. interactions with chatbots or virtual assistants) or they do not even perceive they are dealing with the technology. These interactions are called, respectively, "alterity" and "background" interactions. The nature of these interactions reinforces the argument that AI-nudging can easily be manipulative because it is hidden in two different ways, as I discussed: first, its operation mechanisms are of difficult comprehension; second, AI technologies are increasingly unnoticeable to us. Altogether, I argue, this means that AI-nudging can be more insidious than traditional nudging.

The fact that AI-nudging is arguably more insidious leads to another reason over why AI-nudging presents greater threats to individual autonomy when compared to traditional nudging. Because AI-nudgings are highly hidden and insidious, I argue they are also more pervasive. Let us call a social media user Susan. Kosinski et al (2015, 1037) have found that a computer needs to analyze only 10 of Susan's "likes" in order to portray Susan's personality better than Susan's colleagues. Susan's cohabitants, family members, or spouses are also easily outranked if the computer analyzes 70, 150, or 300 "likes", respectively. It does not come as a surprise, therefore, that Facebook can indeed capture one's emotional state, as discussed earlier.

Moreover, all of our actions in the online world leave a digital trace behind us that unveils our personality, habits, preferences, etc. The data profiles built through data analytics powered by AI contain a varied range of information on us (demographic, financial,

geographical, shopping patterns, political ideologies, etc). Therefore, it is not difficult for the nudger, once in possession of this information, to create highly individualized informational choice architectures, as I suggested earlier in the case of the *Save for Tomorrow* example. In such cases, the nudger is not only nudging one by tapping into one's "System 1" heuristics, but also manipulating one by reaching one's deeper mental state.

The nudging carried out under these circumstances is called by Yeung (2017) "hypernudging", because it creates incredibly individualized environments that are built to match, influence, and manipulate one's most singular idiosyncrasies, leaving individuals a lot more susceptible to manipulation than in the case of typical nudges. Moreover, one's heuristics and vulnerabilities are more exposed in the online world, and AI makes use of this to produce informational choice environments that are increasingly effective in manipulating individuals.

On top of the possibility of "hypernudges", I argue that this pervasiveness of AI-nudging is aggravated by at least two other factors. First, because most AI technologies are deployed by private companies whose best interests are for us to spend as much time as possible on their timelines, many AI models are designed to harness the same level of addiction in us as the one induced by alcohol or drugs (Jones 2020). Moreover, this means that most of the electronic interfaces that mediate our lives are nudging us in the most pervasive ways and will likely continue to do so insofar as it is hard for us to escape them - either because they have addictive effects, or because forgoing such technologies implies forgoing important aspects of contemporary life (e.g. one can live without social networks, but at the cost of missing out some aspects of contemporary social life) or for both reasons concomitantly.

Secondly, AI-nudging can be easily and continuously updated. As an individual leaves more and more digital traces, the choice architecture crafted for this individual can be consistently refined. The more an individual uses streaming services, the better the algorithm of that service will be in offering "optimal" recommendations, because most algorithms are capable of self-learning through feedback loops. Studies on user experience design and privacy have documented the phenomenon of "dark patterns" (Bosch et al 2016). "Dark patterns" refer to the creation of user interfaces (choice architectures) aimed at exploring one's heuristics and vulnerabilities aiming to nudge individuals towards options that they otherwise might not choose (e.g. keep a subscription to a website even when the free trial period is over, or click on a "disguised", camouflaged, advertisement tailor-made to you only because they look like the content of the webpage you are in).

The fact that AI-nudging can be constantly perfected and the fact they are addictive might also arguably create AI-nudging dependency. Especially in the cases in which the AI-nudging does not yield harmful consequences, but instead deliver useful and appreciated results (e.g. the streaming service recommends a movie from a talented new director), individuals might come to appreciate and become dependent on the AI-nudging. This being the case, I argue that, differently from the last chapter - when I defended that it is more worrisome when individuals are nudged towards an end estranged to their expected preferences - AI-nudging is concerning even when it converges to one's preferences.

I claim so because what is at stake here is not purely the outcome yielded by the AI-nudging, but the process that led to this particular outcome. Even if the nudger behind the AI-nudging is a benevolent one, who wants to only pair the nudgee with the best-fitted directions for individuals, the nudger will be making these individuals better off only with respect to the

satisfaction of some of their already existing preferences. There is an aspect of these individual's human flourishing that is being put in jeopardy even in these apparently less harmful cases: the impairment of their individual autonomy. Even if an AI-nudging is well-intended and try to steer individuals towards what it thinks is the best movie for them, these individuals might lose the chance to try and discover something different and pursue their own conception of what a good life is.

Let us consider that a person, Maria, bought tickets for a trip to Paris. Because the website in which she bought the tickets shares Maria's data with, let us say, Google, Maria started to be bombarded with targeted advertising on what to do Paris. The electronic advertises, powered by AI, will likely recommended the most probabilistic results based on the data they have been trained with. For example, when in Paris, most people are likely to visit the *Musée du Louvre* or *Musée D'Orsay*. Because the choice architecture to Maria has been so insisting on convincing her of visiting the Louvre, she might be nudged towards going to this Museum. Maria might arguably enjoy visiting the Louvre, but Maria might also miss the chance to get to know a quirkier, more authentic museum (e.g. the French Postal Museum). AI-nudging, by impairing Maria's individual autonomy through the interference on her decision making, might also compromise Maria's authenticity. Alternatively, without the AI, Maria could have gone to the quirkier museum, but regretted it. But even in that case, and despite the annoyance of losing a couple of Euros and hours, Maria would have been in one way better off: she at least would have learned a lesson that might contribute to her overall personal flourishing.

All in all, the moral risks posed by AI-nudging are also "convergence" risks. As Maria's example shows, Google alone "decides" what is good for Maria. The nudging would arguably

be less worrisome if what Maria decides to do in Paris was influenced by a plurality of influencers. Traditional nudges are relatively pluralistic, meaning that they respond to a variety of interests and actors. AI-nudging, conversely, is quite centralized, insofar as only a few actors control the algorithms. I argue that if it is bad to have one's autonomy impaired, it is even worse when such impairment is enacted by concentrated agents.

In this chapter, being concerned about the morality of AI-nudging, I have argued that AI-nudging can be easily manipulative and even more manipulative than "traditional" nudging. In chapter 1, I defended the view that the necessary condition for a nudge to be morally wrong is that it is manipulative. For a nudge to be manipulative, in turn, I argued it has to be hidden. In this chapter I argued that AI-nudging are especially hidden due to two of its features. First, the sheer operation of an AI-nudging is difficult for humans to process - humans simply lack the cognitive capacity to do so. I mentioned a couple of examples in which AI-nudging has manipulative traits that can be concerning. The examples ranged from less to more concerning cases, varying according to the consequences yielded by the AI-nudging. Yet, as the examples showed, even in the nudges whose consequences seem less harmful, and the nudgee can perceive it - as in the case of a streaming service recommending a movie – the nudgee cannot fully understand how that recommendation came into being. Secondly, AI-nudging is also hidden as far as they are unnoticeable to us. The argument here is that AI-powered technologies are so omnipresent in contemporary life that at some point individuals stop really seeing them for what they are, but only for the means they help to achieve. These two aspects of the "hiddenness" of AI-nudging makes this nudge even more insidiously hidden - and therefore manipulative - than hidden traditional nudges.

This insidious "hiddenness" of AI-nudging leads to another feature of this type of nudge that exacerbates its potential manipulative character - pervasiveness. The choice architectures AI-nudging can build are scarily personalized, meaning they are better suited to exploit even the most intimate of one's heuristics and bypass one's rationality. This pervasiveness is aggravated by other two features: first, some AI-powered technologies have an addictive effect and are, arguably, necessary for a significant part of one's social life - a good one cannot be expected to forgo; second, AI-nudging can be constantly perfected, creating numerous nudging possibilities.

All these features lead to the fact that this type of nudging can make individuals dependent on them. In the previous chapter I argued that the other condition for a nudging to be morally wrong is that, besides being manipulative, it also can lead to an infantilization of the nudgee. The fact that AI-nudging is hidden, pervasive and addictive, might lead one to over-rely on them, in many different aspects of one's life. This might have many different implications: infantilization, loss of one's authenticity, and overall harm to one's personal flourishment. Altogether, these implications point to how AI-nudging can be a threat to individual autonomy and, therefore, can be morally objectionable.

Despite these numerous concerns around the morality of AI-nudging, according to Engelen and Schmidt (2020), "digital choice architecture" is still a "particularly under researched area [in the nudging literature]" (10). I partly agree. I believe that the particularities of AI-nudging have been explored, as indicated in this chapter. The point that preoccupies me now is what can be done, from a regulatory and political point of view, vis-à-vis these threats. The literature on the ethics of nudging is quite keen on addressing the conditions under which it is morally permissible for the state to nudge its citizens (Ivankovic and Engelen 2019). But

the existence of AI-nudges points to another question: what if the nudger is not the state, but actually a private entity, who owns the nudging AI? What is the role of the state vis-a-vis this dynamic? This and other questions shall be addressed in the next chapter.

Chapter 3 – AI-nudging and the State: Regulation and Policy Recommendations

Chapter 2 argued that AI-nudging presents particularities that set it apart from other nudges. Such particularities raise moral concerns that, as I now argue, should trigger state intervention in the form of regulation. I start this chapter by pointing out the reasons why the state ought to regulate AI-nudging, and then I address the current state of affairs regarding digital nudges' regulation, followed by a discussion of how it ideally should be. I first point out that most of the existing legal protection from digital nudges has been grounded in the protection of the rights to freedom of thought or privacy/mental integrity, and not necessarily individual autonomy. If, as laid out in chapter 2, the most pressing objection against AI-nudging is the impairment of one's individual autonomy, then a legal system that places autonomy itself as an overarching good to be protected is desirable. Secondly, I argue that even when certain existing legal dispositions seek to protect autonomy, they usually condition the protection of autonomy on the potential harmful outcomes that the autonomy impairment might yield. If autonomy is morally valuable in itself, as I claimed in chapter 1, then it should be protected because the loss of autonomy is an intrinsic harm, and not (only) because such a loss might yield material harm.

3.1 The State's Duty to Regulate AI-nudging.

The fact that digital nudges, and especially AI-nudging, can so dramatically impair individual autonomy is both a plausible and intuitive reason for the State to intervene in such practices through regulation. Yet, there is another argument behind this logic that strengthens the state's duty to protect individual autonomy, namely the threats to the functioning of a democratic society. By impairing individual autonomy, AI-nudging threatens not only (individual) human flourishing, but it also generates further damages to society at large.

In 1869, John S. Mill wrote the book *On Liberty*, concerned by the "uniformization" of individuals in mass-societies. Such societies, he says, enable informal tools of social vigilance that are different from traditional forms of repression perpetrated by the state, for such tools penetrate "much more deeply into the details of life (...)" (1869, 13). Motivated by this observation, Mill devotes his book to advocating for the preservation of individual liberties. His argument for the preservation of liberties, including the liberty of making up one's own character (which mostly resembles the idea of autonomy), is two-fold: first, individuals should have the chance to develop their characters as independently from undue externalities as possible, as part of their entitlement to human flourishing; secondly, if individuals are deprived of such a right, there is an aggregate loss of creativity and "experiments of [different ways of] living" (Mill 1869, 102), which ultimately might hamper human development. The social consequences of autonomy impairment are more acute in liberal democratic societies, for they lie on the normative assumption that individuals ought, and are capable of, self-ruling. Such a normative principle, in turn, assumes that individuals can set, revise, and pursue their own goals without undue external interference – in other words, it assumes that individuals are autonomous. Tampering with autonomy, I shall argue, also threatens the democratic process.

The fears surrounding informal tools of social vigilance in the 19th century are accentuated by the technologies of the 21st century, both at the individual and social level. As outlined in the previous chapter, AI-nudging is more pervasive and therefore constitutes a greater threat to individual autonomy than other nudges. This logic also manifests itself in the social arena. The 2016 scandal surrounding Cambridge Analytica exemplifies how the AI pollution of individual thinking can yield societal consequences. By gathering around 5.000 "pieces of data" on roughly 220 million American voters, the political consultancy company Cambridge

Analytica built psychological dossiers on these voters and shot highly personalized advertisements at them (Cadwalladr 2016). Such advertisings *hypernudged* the voters towards supporting then-presidential candidate Donald Trump. This case exemplifies how the impairment of one's individual autonomy can yield negative consequences for democracy. To this extent, AI-nudging's indiscriminate and exploitative incursions on individual autonomy is not only an ethical, but also a socio-political issue, and therefore, I argue, both dimensions oblige the state to protect individual autonomy.

Following Mill's liberal thinking, the state may only act against the will of one individual if the aim is to prevent this individual from inflicting harm to others. However, even though Mill claims that individuals whose actions have only private repercussions should not be subjected to state interference – because, according to him, individuals endowed with rational faculties know what is best for them better than anyone else - I believe this is an outdated argument vis-à-vis the current state of technologies, because the latter can undermine the very reasoning behind Mill's argument, namely protecting individual autonomy. Geoffrey Hinton, known as the "godfather of AI", recently claimed that when talking about AI, "you need to imagine something more intelligent than us by the same difference that we are more intelligent than a frog" (Hern 2023, 3). Therefore, even the most intellectually endowed of us can still be susceptible to the manipulative effects of AI-nudging. Additionally, I claim, if individual autonomy is being infringed, then individuals are not even allowed to pursue their own well-being in the first place, for, according to Mill himself, well-being can only be achieved if individuals can choose their own well-being – and for that, one needs to be autonomous.

In my view, any AI that covertly guides the individual in a certain direction, however benign, should be legally impermissible. A libertarian critique that could be raised is that individuals should be allowed to rely on AI guidance if they so wish. By removing the AI-nudging practices, the state is essentially depriving the individual of the autonomy to relinquish autonomy. To reply to this critique, I mobilize the particularities surrounding AI-nudging, and how they can impair individual autonomy in scales previously unforeseen. Therefore, both on the grounds of impairment to (individual) human flourishing and the broader social consequences it might create, I argue that the state has the duty to protect individual autonomy from AI-nudging. How the state ought to do so I discuss in the next section.

3.2 How the State Ought to Regulate AI-nudging

Ahmed Shaheed, United Nations Special Rapporteur on Freedom of Religion or Belief, released in 2021 a Report on the right to Freedom of Thought (FoT), in which he highlighted the threat that emerging technologies pose to this right. Shaheed claimed that emerging technologies question the commonplace belief that thoughts are free before they are expressed, for such technologies pose "dilemmas about how to protect mental privacy, how to protect thoughts from impermissible manipulation and modification (…)." (Shaheed 2021, 26). The report comes in the wake of a wider debate on how human rights can secure individuals' mental integrity vis-à-vis contemporary technological developments and their potential manipulative traits (Ligthart et al 2022). The debate is puzzling because it invites the law to go beyond what is palpable and protect not only material or physical body integrity, but also mental integrity. In this context, scholars (Alegre 2022; Aswad 2020; Bublitz 2020, 2021) and international organizations (United Nations Educational, Scientific and Cultural Organization 2021), have been addressing the topic of how technologies can impair autonomy by focusing on whether the right to FoT offers enough protection to autonomy in this context.

The right to FoT is codified in various legal texts, regionally and globally. Article 18(1) of the International Covenant on Civil and Political Rights (United Nations General Assembly 1966, 10) and Article 9(1) of the European Convention of Human Rights (Council of Europe 1950, 11) both grant the right to "freedom of thought, conscience and religion". In Shaheed's report he takes the definition of FoT to be part of one's "*forum internum* - a person's inner sanctum (mind) where mental faculties are developed, exercised, and defined." (2021, 3). In light of such a well-established human right, the academic debate is, thus, around whether FoT satisfactorily protects the challenges posed by current technologies to individual's autonomy, or if new human-rights are needed in order to ensure proper protection of autonomy.

Amidst this debate, I argue that the right to FoT, and the way it has been historically legally interpreted, is not enough to protect autonomy vis-à-vis manipulative technologies, such as AI-nudging. I defend the view that, instead, regulation aimed at protecting individual autonomy, in the context of technology regulation, should be grounded on autonomy itself, or correlated with terms such as the right to "mental integrity" or "self-determination", which capture more precisely what individual autonomy is, and which is in line with what the European Union, under its European Data Strategy, has been increasingly proposing in its regulatory pieces[2]. Overall, I argue that in the context of AI-nudging, individual autonomy, and not FoT, is a good to be intrinsically protected by the law.

[2] I opted to focus my analysis on the European regulatory landscape because, historically, the European Union is pioneer when it comes to tech policy. The EU was the first jurisdiction, for example, to set horizontal regulatory standards to regulate AI, arguably exporting its regulatory models to the rest of the world.

In the realm of philosophy, Mill connects FoT to freedom of expression (1869, 31-99). In this sense, the moral value of FoT is correlated with the fact that FoT enables one to search for truth. Therefore, the threats posed to FoT have been traditionally understood to be indirect only: because until recently the understanding was that it was not possible to interfere with one's thoughts *directly*, it was assumed that FoT could be violated only through *indirect* facts, such as censorship. Moreover, because until recently technologies such as the one's developed by Elon Musk's Neuralink – which literally connects human brains to machines by chip implantations - were inconceivable, thoughts were considered to be possible to harm only indirectly. Therefore, when it comes to court cases, FoT has not, in general, been interpreted as being violated not because one's thoughts have been infringed themselves, but rather because the expression of one's thoughts (speech) has been censured. According to Moore (2022), by "not perceiving any need to invoke Freedom of Thought itself, judges neglected it, failing to give it in practice any content of its own" (516). The United Nations Human Rights Committee has oftentimes chosen to not analyze cases through the lens of FoT even when the plaintiff has claimed FoT violations.[3] The committee has rather relied its deliberations on other human rights – freedom of expression most notably.

Indeed, the very definition of what FoT and what it encompasses it still debated. As Shaheed writes, different stakeholders (legal practitioners, linguists, philosophers, etc) attribute different, and oftentimes conflicting, definitions of what a "thought" is, rendering FoT a right that "not only lacks legal precision but also scientific and philosophical consensus" (2001, 5). Moreover, even though the right is extensively codified in different legal systems and treaties,

[3] See United Nations Human Rights Committee (2005), Communication No. CCPR/C/84/D/1119/2002, 7-8, § 7.4.

its exact meaning and application remains imprecise. I believe this uncertainty around FoT, with blurred lines between FoT and freedom of expression, is problematic in the context of manipulative technologies, insofar as, according to Shaheed, the UN Rapporteur, current technologies can interfere with one's thoughts before they are expressed, so the protection of one's mental processes should rely on potential infringements to the individual autonomy itself.

In this sense, while FoT is traditionally a prominent right, it is also a "tainted" right, for its interpretation is oftentimes associated with freedom of expression. Such "contamination", I argue, makes FoT a right that is not suitable for properly protecting individuals from the mental incursions perpetrated by AI-nudging, because AI-nudging threatens not the expression of one's thoughts, but rather one's ability to autonomously think in the first place. Moreover, the right to FoT, which has been historically interpreted and associated with freedom of expression, would constitute an adequate protection from AI-nudging if AI-nudging threatened the expression of one's thoughts. But that is not the case, for what AI-nudging actually threatens is one's capacity to set, revise and pursue one's own goals – it threatens one's individual autonomy primarily.

Recently, in light of the latest technological advancements, some scholars have sought to interpret FoT in a broader sense. Ligthart et al (2022, 2) identify three constitutive elements of FoT, one of them seeking to ensure that thoughts cannot be "impermissibly altered". Such definition, I note, does not necessarily offer protection to individual autonomy, for it places it as only one, out of the many other factors FoT can encompass. In Ligthart's definition, for example, FoT also includes the notion that persons should not be "forced to reveal their thoughts", nor "sanctioned" for their thoughts. The right to autonomously thinking is only one, out of many, other rights, and even in this definition, the legal consequences associated with

one's thoughts are pinned on the expression/non-expression of those thoughts – not on the ability to think itself. In order words, the definition of FoT remains blurry and tainted (even if it is no longer tainted by its connection to the right to freedom of expression). If, as discussed in the last chapter, the nudges perpetrated in digital choice architectures are more morally worrisome than traditional nudges, when it comes to autonomy impairment, then a legal system with rights aimed at protecting precisely individual autonomy from manipulative technologies is desirable.

This is in line with what legal scholars have suggested: Susie Alegre (2022) acknowledges the limitations of the right to FoT in the current context of digital technologies and proposes that, if FoT is to constitute an adequate response to technology-powered incursions in individual's minds, then it needs to be significantly reconceptualized; Andorno and Ienca (2017) propose the creation of "neurorights" encompassing the right to cognitive liberty, mental privacy, mental integrity, and psychological continuity.; in the middle ground, Jan C. Bublitz (2020, 2021) suggests both the reinterpretation of the right to FoT, but also the creation of rights aimed at protecting mental self-determination. Hertz (2023) argues that such new rights would not necessarily bring any extra layer of protection, and that the key to protect one's individual autonomy lies in legal reinterpretations of FoT . In this debate, I position myself closer to those who advocate for the creation of rights that are tailor-made to the current issues raised by technological impairment to individual autonomy.

One could argue that creating new human rights might lead to a situation in which the extensive number of human rights generates unattainable compliance requirements, potentially even decreasing legal clarity (Alston 1984; Baxi 2001). I believe such critique is outweighed by the perspective that we are currently confronted with new human rights issues (i.e.

autonomy impairment caused by manipulative technologies) that still remain largely unattended, either because existing rights, such as FoT, fail to capture all the particularities arisen by new technologies, or because such issues are simply not yet fully acknowledge. In 2019, the Commissioner for Human Rights in the Council of Europe published a Recommendation called "Unboxing Artificial Intelligence: 10 steps to protect human rights" without any mention to the right to FoT or the need to protect individual autonomy, mental integrity, or mental self-determination from AI-powered technologies (Council of Europe 2019). I support the idea, thus, that the creation of new rights would fulfill the need to draw special attention to the technological threats to individual autonomy, a phenomenon that is arguably being overlooked.

Additionally, the need to protect individual autonomy is not unfamiliar to the legal domain. The concept of autonomy itself is no legal novelty. Hans Kelsen (1950) actually places autonomy as a vital feature of any legal system for two main reasons. First, it is only because individuals are assumed to be rational and autonomous that legal norms exist and that one can reasonably expect that such norms will be followed. Secondly, this same assumption is what substantiates the idea that legal culpability can be inputted to individuals. In contractual law, for example, the assumption that individuals enjoy freedom and autonomy is likewise essential for the application of legal provisions. When it comes to technology regulation, the concept of autonomy cannot simply be *ipsis literis* borrowed from other legal areas such as contractual law. The regulation of technologies holds its own particularities, for technologies present their own moral concerns, as exemplified in the last chapter. Yet, this does not mean that the legal system must remain as it is. Individual autonomy can be better protected from AI-nudging, and any other technology-powered manipulation, if there are rights focused on protecting autonomy itself. The European regulators have shown that this is possible, as I now show.

Since the General Data Protection Regulation (GDPR) - the first comprehensive EU regulation in the realm of technology regulation – entered into force in 2018, the European regulators have been drawing increasing attention to the need to protect individual autonomy. The intent to protect autonomy was started covertly, hidden behind the intent to protect privacy, and as the years passed it started to become more apparent, culminating in legal provisions explicitly stating that their very *raison d'être* is to protect autonomy. Such legal provisions are concentrated under some of the legal instruments integrating the European Strategy for Data, a policy launched by the European Commission in 2020 and aimed at creating a single market for data commercialization within the EU, while strengthening the pursuit of "European values" in the digital world (European Commission 2020, 4-5).

One of these legal instruments, for example, the Data Act draft, published in 2022, aims to bring legal clarity as to how value can be created out of both personal and non-personal data, i.e., how these data can be processed. It focuses foremost on those companies which "recycle" data collected firstly by another data processor. Under its Article 6(2), subparagraph (a), the Data Act establishes that such third-party companies shall not "coerce, deceive or manipulate the user in any way, by subverting or impairing the autonomy, decision-making or choices of the user, including by means of a digital interface with the user;" (European Commission 2022, 43). I argue that by being explicitly mentioned in the regulation, individual autonomy would benefit from direct protection, and not as an extension to safeguards to privacy that sought to protect one's mental integrity.

In the European context, regulators have indirectly protected individual autonomy while seeking to safeguard privacy and mental integrity since the enactment of the GDPR. This

is not necessarily condemnable, but rather a natural phenomenon, given that autonomy, privacy, and mental integrity rights are closely intertwined. Therefore, it is no surprise that the GDPR, a regulation aimed at precisely safeguarding privacy in the digital realm, taps on, and indirectly protects, individual autonomy, through its aims to secure mental integrity. According to Gartner (2022), the "lawfulness, fairness, and transparency" guiding principle for data processing, established under Article 5(1), subparagraph (a) of the GDPR (European Parliament and Council 2016, 35), was interpreted by the European Data Protection Board (EDPB) as not being compatible with practices associated with autonomy-constraining nudges (EDPB 2020a). In other words, the lawfulness principle seems to, under certain circumstances, offer protection to individual autonomy.

Additionally, the Article 6(1), subparagraph (a), of the GDPR (European Parliament and Council 2016, 36) establishes the conditions in which data processing is lawful - the necessity for the data subject's consent being the most prominent of this condition. This disposition constitutes a turn towards autonomy protection to the extent that, as pointed out by the EDPB, "consent will not be free in cases where there is any element of compulsion, pressure, or inability to exercise free will" (EDPB 2020b, 9). Moreover, the processing of any individuals' personal data is not lawful if individuals do not benefit from full autonomy at the time they consented to their data being processed. The jurisprudence from the European Court of Justice, I add, reinforces this understanding. For example, in 2019, the Court ruled, that pre-ticked consent boxes are not sufficient to establish that an internet user has consented to the processing of their personal data, for they do not reflect one autonomous act of giving consent.[4]

[4] See C-673/17, Bundesverband der Verbraucherzentralen und Verbraucherverbände — Verbraucherzentrale Bundesverband eV *v* Planet49 GmbH., 2019, E.C.J, ECLI:EU:C:2019:801.

As the years went by, the protection of autonomy became more evident and independent of privacy safeguards. According to Gartner (2022), in the very first (leaked) version of the EU Artificial Intelligence Act prohibited AI systems "manipulates human behaviour, opinions or decisions through choice architectures or other elements of user interfaces, causing a person to behave, form an opinion or take a decision to their detriment" (468). More recently, in 2022, the Digital Services Act (DSA) – another legal instrument under the umbrella policy European Strategy for Data - entered into force echoing the Data Act's call for autonomy protection by explicitly mentioning the term "autonomy". Amongst all the dispositions on the Act, I highlight how the DSA also seeks to protect individual autonomy from another especially worrisome future of AI-nudgings mentioned in chapter 2 - the so-called "dark patterns". Under its Recital 67, the DSA establishes that:

> "Providers of online platforms should therefore be prohibited from deceiving or nudging recipients of the service and from distorting or impairing the autonomy, decision-making, or choice of the recipients of the service via the structure, design or functionalities of an online interface or a part thereof. This should include, but not be limited to, exploitative design choices to direct the recipient to actions that benefit the provider of online platforms, but which may not be in the recipients' interests, presenting choices in a non-neutral manner (…)" (European Parliament and Council 2022, 18).

The DSA and the DA indicate the latest step towards the aim to protect individual autonomy from technologically-driven mental incursions on the grounds of autonomy itself, initiated by the GDPR and modestly advanced in the EU AI Act. In both the DSA and DA, autonomy is not being protected indirectly on the grounds of FoT or privacy, but rather it is mentioned explicitly. The dispositions themselves, in both cases, were framed as to legally capture as much as possible the empirical features of digital nudges – especially those powered by AI – that make them a threat to individual autonomy. The Articles are careful and accurate enough to encompass, and seek to mitigate, all the features I laid out in chapter 2 as being constitutive of morally wrongful/manipulative AI-nudging: the hiddenness, pervasiveness, and excessive individualization of the choice architectures. This, I argue, shows that the European regulators are perspicacious about the particular moral hazards posed by AI-nudging. The

acknowledgement of such particularities is shown in the commitment to develop the legal system by protecting autonomy as something that, as argued in chapter 1, is morally valuable in itself. The regulators have promisingly evolved to understand that individual autonomy is directly under threat, and therefore it is not sufficiently protected by proxy-rights.

Pitfalls to be avoided, I argue, include the temptation to protect individual autonomy conditioned to the potential harms that the loss of autonomy might yield. The version of the EU AI Act currently being voted largely differs from the first draft I transcribed previously. Whilst that leaked version prohibited "AI systems that manipulate human behavior, opinions or decisions (…)" (Gartner 2022, 468), indicating the understanding that the impairment of autonomy is morally problematic in itself, the current version of the EU AI Act conditions the wrongness of the AI-nudging to the potential psychological or physical harms the autonomy impairment might cause. Article 5(1), subparagraph (a), establishes that AI is prohibited if it "(…) deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behavior in a manner that causes or is likely to cause that person or another person physical or psychological harm;" (European Commission 2021, 43). The attempt to condition the harm to autonomy to material consequences is understandable: the law is used to protect what is palpable, and when what is at stake is a highly abstract moral concern such as autonomy, then the law is challenged.

I argue that such attempt, despite being understandable, ought to be avoided, or else the right to individual autonomy might fall under the same trap as the right to FoT, a right whose content is still highly debatable precisely because it was not interpreted according to its own meaning, but only in light of harms to freedom of expression. Thus, I argue, the potential failure to give the protection to autonomy its own content is not only a failure to grasp that autonomy

is morally valuable in itself, but might also render such protection inefficient. In light of the desirability to avoid this result, the attempt to protect individual autonomy from AI-nudging on the grounds of autonomy - and not FoT and privacy - is already an achievement. But such accomplishment can be furthered if it is accompanied by policies that seek to bring more concreteness to the protection of autonomy, without conditioning it on the material consequences that infringement of autonomy has on individuals.

In the last chapter I argued that one of the conditions that makes a nudge (AI-powered or not) morally wrongful is the manipulativeness of the nudge. The manipulation, in turn, is said to happen if the nudge is hidden. In the case of AI-nudging, I argued this hiddenness is exacerbated, insofar as humans cannot understand the nudge they are subjected to, and because one oftentimes barely notices the technologies one is nudged by. In this sense, it is possible to conceive policies that seek to address this precise issue. Companies could be required to disclose, in a comprehensible way, the basic operational features behind their nudging-algorithms, provided their intellectual property rights are safeguarded. If such a balance between the informational knowledge of the algorithms' functioning and company's business secret could be achieved, then a scenario in which the AI-nudgees are not manipulative is more feasible. Similarly, policies aimed at making users more aware of the technologies they interact with are likewise welcomed. As discussed in chapter 2, whereas in the last year User Experience professionals have sought to make interactions with technology as seamless as possible, maybe some frictions are needed in order to make individuals more conscious about the technologies they use and the decisions they are nudged to. Choice architectures could be built so as to foster some hesitation before users take decisions, for example, presenting pop-ups questioning the users if they really want to proceed with that shopping, that message, etc.

All these policies, I acknowledge, would face criticism from the tech sector, potentially rendering the policies politically unfeasible. For example, the significant difference spotted in the provisions of the first draft of the EU AI Act and its current version is due to, in large part, political pressure from the tech sector and subsequent political surrendering by the side of the legislators. A report produced by Corporate Europe Observatory found out that big tech corporations have poured billionaire resources into Brussels lobbying in order to "water down the EU AI Act" (Schyns 2023, 32), especially when it comes to the regulation of General Purpose AI (GPAI). Yet, I argue that such measures, if accompanied by the current trend of explicitly protecting autonomy, as indicated by the DSA and Data Act, might constitute, altogether, a good example of how to translate the autonomy concerns raised by AI into concrete policy and regulatory propositions.

Conclusions

At the time I write this dissertation, the world is grappling to understand the real economic and social impacts associated with AI. Amongst the many concerns that could be risen, I focused in this dissertation on the ethical impact that AI has on human decision-making, and how such concerns can be properly mitigated by regulation and policies. I opted for this focus because the threats AI-nudging presents to human autonomy are still unforeseen, and since individual autonomy is, arguably, the key to human flourishing and to the attainment to a handful of other rights, the issue demands academic and political attention.

My main argument is that AI-nudging possesses features that makes AI incursions into human thinking far more ethically concerning than those perpetrated by traditional nudgings. This is because AI-nudging is more manipulative than traditional nudging, for it can be especially hidden – humans can hardly fully understand it, and AI-driven technologies are increasingly unnoticeable. By being unnoticeable, I argued, such nudges are more insidious and pervasive – two features that accentuate the exploitative character of AI-nudging, especially when considering that any nudge performed by an AI is highly individualized, turning the nudgee particularly exposed and prone to being manipulated. Additionally, such nudges can be "addictive," going around individual's rationality indeterminately, raising "convergence" concerns. For all these reasons, I argue that AI-nudging not only fulfills the two conditions I stipulated in chapter I for a nudge to be morally wrongful, but it exacerbates them, impairing individual autonomy.

The main conclusion I take from this is that the state has the duty to protect individual autonomy, and the way the state ought to do so is by forging a legal system that protects individual autonomy in itself, avoiding the pitfalls of relying on rights that are already well-

established, such as freedom of thought or privacy, but which do not entirely address all the particularities of AI-nudging. The driving distinction behind my analysis is that any legal system that seeks to adequately safeguard individual autonomy must move away from a legal framework that attributes salience to "harm" (e.g. adverse material consequences that may stem from AI-nudging) towards laws that protect individual autonomy irrespective of consequences and manifestations. Whereas my conclusion bears the challenge of moving the law from the material to the perceptive level, I provide evidence that the tech legal framework of the European Union, while still in its infancy, this legal framework is already being built up. I argue this is the key for the moral concerns risen in chapter 2 to be adequately addressed as to ensure individual autonomy protection in the age of AI-nudging.

Bibliography

Alegre, Susie. 2022. *Freedom to Think: The Long Struggle to Liberate Our Minds*. London: Atlantic Books.

Alston, Philip. 1984. "Conjuring up new human rights: A proposal for quality control". *The American Journal of International Law* 78 (3): 607-621. https://doi.org/10.2307/2202599 .

Andorno, Roberto and Marcello Ienca. 2017. "Towards new human rights in the age of neuroscience and neurotechnology". *Life Sciences, Society and Policy* 13 (5): 1-27. http://dx.doi.org/10.1186/s40504-017-0050 .

Aswad, Evelyn M. 2020. "Losing the Freedom to Be Human". *Columbia Human Rights Law Review* 52(1): 306 - 371. https://ssrn.com/abstract=3635701 .

Atillah, Imane El. 2023. "Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change." *Euronews*, March 31st, 2023. https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate- .

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact". *California Law Review* 104(3): 671-732. https://doi.org/10.15779/Z38BG31 .

Baxi, Upendra. 2001. "Too Many, or Too Few, Human Rights". *Human Rights Law Review* 1(1): 1–10. https://doi.org/10.1093/hrlr/1.1.1 .

Benartzi, Shlomo and Richard H. Thaler. 2004. "Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving." *Journal of Political Economy* 112 (1): 164-187. https://doi.org/10.1086/380085 .

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity Press.

Berlin, Isaiah. 1997. *The Proper Study of Mankind*. London: Chatto & Windus.

Bosch, Christoph, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. "Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns." *Proceedings on Privacy Enhancing Technologies* (4) 237-254. https://doi.org/10.1515/popets-2016-0038 .

Bovens, Luc. 2009. "The ethics of nudge". In *Preference change: Approaches from philosophy, economics and psychology*, edited by Till Grüne-Yanoff and Sven O. Hansson, 207–219. Berlin and New York: Springer Science & Business Media.

Bublitz, Jan C. 2020. “The nascent right to psychological integrity and mental self-determination.” In *The Cambridge handbook of New Human Rights*, edited by Andreas von Arnauld, Kerstin von der Decken and Mart Susi, 387-403. Cambridge: Cambridge University Press.

——. 2021. “Freedom of thought as an international human right: Elements of a theory of a living right.” In *The Law and Ethics of Freedom of Thought, Volume 1 – Neuroscience, Autonomy, and Individual Rights*, edited by Jan Christoph Bublitz and Marc Jonathan Blitz, 49-101. Cham: Palgrave Macmillan.

Buss, Sarah and Andrea Westlund. 2018. "Personal Autonomy". *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/ .

Cadwalladr, Carole. 2016. “Google, democracy and the truth about internet search.” *The Guardian*, December 4th, 2016. https://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook .

Chakrabortty, Aditya. 2008. “From Obama to Cameron, why do so many politicians want a piece of Richard Thaler?”, *The Guardian*, July 12, 2008. https://www.theguardian.com/politics/2008/jul/12/economy.conservatives.

Clarke, Roger. 1988. “Information technology and dataveillance”. *Communications of the ACM* 31(5): 498-512. https://doi.org/10.1145/42411.42413.

Coeckelbergh, Mark. 2022. *The Political Philosophy of AI*. Cambridge and Medford: Polity Press.

Cohen, Julie. E. 2012. *Configuring the Networked Self*. New Haven: Yale University Press.

Council of Europe. 1950. *European Convention of Human Rights*. Strasbourg: Council of Europe. https://www.echr.coe.int/documents/convention_eng.pdf .

——. 2019. *Unboxing Artificial Intelligence: 10 steps to protect Human Rights*. Strasbourg: Council of Europe. https://www.coe.int/en/web/commissioner/-/unboxing-artificial-intelligence-10-steps-to-protect-human-rights .

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Dworkin, Gerald. 1988. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press. https://books.google.at/books/about/The_Theory_and_Practice_of_Autonomy.html?id=wcFGH-zIyGgC&redir_esc=y .

Engelen, Bart and Andreas T. Schmidt. 2020. “The Ethics of Nudging: An Overview”. *Philosophy Compass* 15(4): 1-13. https://doi.org/10.1111/phc3.12658 .

European Commission. 2020. *Communication from the European Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the European Strategy for Data.* Brussels: European Commission, (EU) COM(2020) 66 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066 .

——. 2021. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.* Brussels: European Commission, COM/2021/206 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 .

——. 2022. *Legislative Proposal of the European Commission for a Regulation of the European Parliament and of the Council on Harmonised Rules on Fair Access to and use of data (Data Act).* Brussels: European Commission, (EU) COM(2022) 68. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN .

European Data Protection Board (EDPB). 2020a. *Guidelines 4/2019 on Data Protection by Design and by Default.* Brussels: European Data Protection Board. https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf .

——. 2020b. *Guidelines 05/2020 on consent under Regulation 2016/679.* Brussels: European Data Protection Board. https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf .

European Parliament and the European Council. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* Brussels: Official Journal of the European Union, (EU) 2016/679. https://eur-lex.europa.eu/eli/reg/2016/679/oj .

——. 2022. *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).* Brussels: Official Journal of the European Union, (EU) 2022/2065. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065 .

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York: Picador, St Martin's Press.

Frankfurt, Harry G. 1988. *The Importance of What We Care About.* Cambridge: Cambridge University Press.

Gartner, Maximilian. 2022. "Regulatory Acknowledgment of Individual Autonomy in European Digital Legislation: From Meta-Principle to Explicit Protection in the Data Act." *European Data Protection Law Review* 4(8): 462 - 473. https://doi.org/10.21552/edpl/2022/4/6 .

Grüne-Yanoff, Till. 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 .

Halpern, David. 2015. *Inside the nudge unit: How small changes can make a big difference*. London: WH Allen.

Hausman, Daniel. M and Brynn Welch. 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 .

Heidegger, Martin.1962. *Being and Time*. Translated by John Macquaire and Edward Robinson. Oxford: Blackwell Publishers.

Hern, Alex. 2023. “Interview - ‘We’ve discovered the secret of immortality. The bad news is it’s not for us’: why the godfather of AI fears for humanity.” *The Guardian*, May 5th, 2023. https://www.theguardian.com/technology/2023/may/05/geoffrey-hinton-godfather-of-ai-fears-for-humanity .

Hertz, Nora. 2023. “Neurorights: Do We Need New Human Rights? A Reconsideration of the Right to Freedom of Thought.” *Neuroethics* 16(5): 1-15. https://doi.org/10.1007/s12152-022-09511-0 .

Ihde, Don. 1990. *Technology and life world: from garden to Earth*. Bloomington: Indiana University Press.

Ivankovic, Viktor and Bart Engelen. 2019. “Nudging, transparency, and watchfulness.” *Social Theory and Practice* 45(1), 43–74. https://doi.org/10.5840/soctheorpract20191751 .

Jones, Kate. 2020. "Mental Autonomy Must be Preserved as Tech Advances." *Catham House*, December 17, 2020. https://www.chathamhouse.org/2020/12/mental-autonomy-must-be-preserved-tech-advances .

Kahneman, Daniel. 2011. *Thinking, fast and slow*. London : Allen Lane ; Penguin Books.

Kelsen, Hans. 1950. “Causality and imputation.” *Ethics* 61 (1):1-11. https://www.jstor.org/stable/2379043 .

Kosinski, Michal, David Stillwell and Wu Youyou. 2015. “Computer-based personality judgments are more accurate than those made by humans”. *Proceedings of the National Academy of Sciences* 112(4): 1036-1040. https://doi.org/10.1073/pnas.1418680112 .

Leggett, Will. 2014. “The Politics of Behaviour Change: Nudge, Neoliberalism and the State.” *Policy & Politics* 42(1): 3–19. https://doi.org/10.1332/030557312X655576 .

Levin, Sam T. 2017. “Facebook told advertisers it can identify teens feeling 'insecure' and 'worthless'”. *The Guardian*, May 1st, 2017.

https://www.theguardian.com/technology/2017/may/01/facebook-advertising-data-insecure-teens .

Ligthart, Sjors, Christoph Bublitz, Thomas Douglas, Lisa Forsberg and Gerben Meynen. 2022. “Rethinking the Right to Freedom of Thought: A Multidisciplinary Analysis.” *Human Rights Law Review* 22(4): 1-14. https://doi.org/10.1093/hrlr/ngac028 .

MacKay, Douglas and Alexandra Robinson. 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 .

Macleod, Christopher. 2020. "John Stuart Mill”. *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/mill/#toc .

Mill, John Stuart. 1869. *On Liberty*. London: Longmans, Green, Reader, and Dyer. https://en.wikisource.org/wiki/On_Liberty .

Moore, Marcus. 2022. “Freedom of thought at the ethical frontier of law & science.” *Ethics & Behavior* 32(6): 510-531. https://doi.org/10.1080/10508422.2021.1928500 .

Nissenbaum, Helen, Beate Roessler and Daniel Susser. 2019. “Technology, Autonomy and Manipulation.” *Internet Policy Review* 8(2):1-22. https://doi.org/10.14763/2019.2.1410 .

Pasquale, Frank. 2015. *The Black Box Society; The Secret Algorithms that Control Money and Information*. Cambridge: Harvard University Press. https://www.jstor.org/stable/j.ctt13x0hch .

Peeters, Rik. 2019. “Manufacturing Responsibility: The Governmentality of Behavioural Power in Social Policies.” *Social Policy and Society* 18(1): 51-65. https://doi.org/10.1017/S147474641700046X .

Raz, Joseph. 1986. *The Morality of Freedom*. Oxford: Oxford University Press.

Schyns, Camille. 2023. *The Lobbying Ghost in the Machine*. Brussels: Corporate Europe Observatory. https://corporateeurope.org/sites/default/files/2023-02/The%20Lobbying%20Ghost%20in%20the%20Machine.pdf .

Shaheed, Ahmed. 2021. A/76/380: Interim report of the Special Rapporteur on freedom of religion or belief, Ahmed Shaheed: Freedom of thought. Geneva: United Nations Human Rights Office of the High Commissioner. https://www.ohchr.org/en/documents/thematic-reports/a76380-interim-report-special-rapporteur-freedom-religion-or-belief .

Simon, Herbert A. 1955. “A behavioral model of rational choice.” *The Quarterly Journal of Economics*, 1(69): 99–118. https://doi.org/10.2307/1884852 .

Solove, Daniel J. 2004. *The Digital Person: Technology and Privacy in the Information Age*. London and New York: New York University Press. https://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=2501&context=faculty_publications .

Sunstein, Cass R., and Richard H. Thaler. 2008. *Nudge: improving decisions about health, wealth, and happiness*. New Haven: Yale University Press.

Susser, Daniel. 2017. "Transparent Media and the Development of Digital Habits". In *Postphenomenology and Media: Essays in Human-Media-World Relations*, edited by Yoni Van den Eede, Stacy O'Neal Irwin and Galit Wellner, 27-44. New York: Lexington Books.

Tversky, Amos and Daniel Kahneman. 1974. "Judgment under uncertainty: heuristics and biases." *Science* 185(4157):1124–1131. https://doi.org/10.1126/science.185.4157.1124 .

United Nations Educational, Scientific and Cultural Organization. 2021. *Report of the International Bioethics Committee of UNESCO (IBC) on the Ethical Issues of Neurotechnology*. Paris: UNESCO, SHS/BIO/IBC-28/2021/3. https://unesdoc.unesco.org/ark:/48223/pf0000378724 .

United Nations General Assembly. 1966. *International Covenant on Civil and Political Rights (ICCPR)*. New York: United Nations General Assembly. https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights .

United Nations Human Rights Committee. 2005. Communication CCPR/C/84/D/1119/2002. Geneva: United Nations Human Rights Committee. https://digitallibrary.un.org/record/560805 .

Verbeek, Peter P. 2005. *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Translated by Robert P. Crease. Pennsylvania: Penn State University Press. https://www.jstor.org/stable/10.5325/j.ctv14gp4w7 .

Yeung, Karen. 2017. "'Hypernudge': Big Data as a Mode of Regulation by Design". *Information, Communication & Society* 20 (1): 118–36. https://doi.org/10.1080/1369118X.2016.1186713 .