AI & SOCIETY
https://doi.org/10.1007/s00146-022-01486-z

ORIGINAL PAPER

The autonomous choice architect

Stuart Mills[1] · **Henrik Skaug Sætra**[2]

Received: 11 August 2021 / Accepted: 28 April 2022

Abstract
Choice architecture describes the environment in which choices are presented to decision-makers. In recent years, public and private actors have looked at choice architecture with great interest as they seek to influence human behaviour. These actors are typically called choice architects. Increasingly, however, this role of architecting choice is not performed by a human choice architect, but an algorithm or artificial intelligence, powered by a stream of Big Data and infused with an objective it has been programmed to maximise. We call this entity the autonomous choice architect. In this paper, we present an account of why artificial intelligence can fulfil the role of a choice architect and why this creates problems of transparency, responsibility and accountability for nudges. We argue that choice architects, be them autonomous computational systems or human-beings, at a most basic level select, from a range of designs, the design which is most likely to maximise a pre-determined objective. We then proceed to argue that, given the growing demand for targeted, personalised choice architecture and for faster, dynamic reconfigurations of choice architecture, as well as the ever-expanding pool of data from which feedback can be drawn, the role of the human choice architect is increasingly obscured behind algorithmic, artificially intelligent systems. We provide a discussion of the implications of autonomous choice architects, focusing on the importance of the humans who programme these systems, ultimately arguing that despite technological advances, the responsibility of choice architecture and influence remains firmly one human beings must bear.

Keywords Choice architecture · Artificial intelligence · Nudge · Behavioural science · Big data

JEL Classification D9 · D91

1 Introduction

Choice architecture describes the environment in which choices are presented to decision-makers, or the design of those presentations (Hausman and Welch 2010; Thaler et al. 2012). The logical masters of choice architecture, so-called choice *architects*, are said to be tasked with the meaningful architecting of choices so as to influence the actions of decision-makers in an intentional way, without restricting options or significantly changing economic incentives (Thaler and Sunstein 2008). The notion of choice architecture emerges as a necessary component of *nudge theory*, a subset of behavioural science which seeks to meaningfully redesign choice architecture so as to influence decision-makers based on their behavioural biases, without restricting freedom of choice (Sunstein 2014).

Since its popular inception, most nudging and the architecting of choices has been performed by humans, usually in teams arranged as private consultancies or, more frequently, appendages of government (Sanders et al. 2018; Thaler and Sunstein, 2008). Increasingly, however, the architecting of choices is an automated activity, seemingly either devoid of a human choice architect, or devoid of the oversight of a human choice architect (Jameson et al. 2013; Mele et al. 2021;

The authors are grateful to Richard Whittle for helpful comments on an earlier draft of this paper, and to Liam Delaney for allowing an early draft of this paper to be presented at the LSE's *Behavioural Science and the Wider World* seminar series. All errors are the authors' own. We would also like to extend our thanks to the anonymous reviewers, whose comments have greatly improved this article.

✉ Stuart Mills
s.mills3@lse.ac.uk

[1] Department of Psychological and Behavioural Science, London School of Economics and Political Science, London, UK

[2] Faculty of Computer Sciences, Engineering and Economics, Østfold University College, 1757 Halden, Norway

Published online: 22 June 2022

Weinmann et al. 2016; Yeung 2017). Take, for instance, the Facebook News Feed algorithm, which will curate, on a daily basis, around 300 posts to appear to the individual Facebook user, out of an average of around 1500 possible posts (Luckerson 2015). This is nudging via the meaningful design of choice architecture (Johnson et al. 2012); per Thaler and Sunstein (2008), no choices are mandated or banned (a user can always manually go onto pages and see what's changed), posts are curated because humans are assumed to have bounded cognitive processes (e.g., bias; Simon 1955), and the posts which are selected are those which are predicted to be most interesting to the user (Luckerson 2015).

Insofar as Facebook and similar entities (to name a few: Google and its subsidiary services such as YouTube and Google Maps; Amazon; Microsoft and its subsidiary services such as LinkedIn; Facebook subsidiary services such as Instagram) *architect choices*, these entities—or rather, the algorithms they design and implement—would seem to function as *autonomous choice architects* (Johnson 2021; Johnson et al. 2012; Lavi 2017).

In this paper, we will explore the idea of autonomous choice architects, relating the concept to machine learning and AI. The notion that AI is now used to influence individual behaviour in ways both compatible to the concept of nudging and in other ways is not new (Helbing 2015; Jameson et al. 2013; Yeung 2017). However, exactly how AI may play the role of a choice architect is as of yet relatively underexplored (Mele et al. 2021), which is why we go into some detail regarding what nudging is, what AI is, and why we argue that AI can in fact be a choice architect. Having established this, we provide a novel foundation for discussing AI and nudging by avoiding imprecise and arguably erroneous accounts of how these concepts are related.

We define some terms in relation to these ideas in Sect. 2. Equipped with this background, we advance a perspective which regards choice architects as *selectors*, borrowing from Yeung's (2017) argument that various technologies, such as machine learning, AI and algorithms more broadly, merely enable "selection optimisation" (p. 121) when it comes to nudging and choice architecture. Indeed, we argue that, at its core, to *architect a choice* is to *select from a set of possible designs that which is predicted to be most effective at influencing the decision of a decision-maker in accordance with a pre-determined objective*. This perspective may evoke criticism from various parties. To be clear, our perspective concerns the selection of choice architecture (i.e., nudging) from a set of known choice architectural techniques. Thus, we do not necessarily dismiss the creative and experimental component of designing nudges, which at present remains the reserve of *human* choice architects; we simply argue that once various components are designed, the process of changing choice architecture (i.e., nudging) is one of selection from this set of designs, a task which can be performed algorithmically or via an AI system. Furthermore, our discussion concerns autonomous systems insofar as they *architect choices*; we will not generally engage with criticisms of these systems beyond those which arise at this intersection (Susser et al. 2019; Zarsky 2019).

We regard our selection perspective on choice architecture to also be quite intuitive. Consider the popular example of choice architecture where the default option given to a decision-maker is changed. People tend to choose whatever option is given as the default option (Jachimowicz et al. 2019). One striking example of this is given by Johnson and Goldstein (2004), who find that defaulting people into being organ donors, rather than defaulting them into *not* being donors, leads to around 98–99% of people choosing to become donors, across several European countries. Jachimowicz et al. (2019) offer many further examples in their review of the default option literature, most less dramatic than Johnson and Goldstein (2004) in effectiveness, but still generally demonstrably effective. From a very *abstract, theoretical perspective*, choice architects who are tasked with setting a default are simply tasked with selecting which option from a set $\{A, B, C \ldots n\}$ should be set as the default (though this is not necessarily a simple task in actuality; Johnson et al. 2012). Traditionally understood, choice architects cannot expand or reduce the number of items in this set, and ultimately will decide to set as the default whatever option is deemed most effective, however this is determined.[1]

The perspective of the *choice architect as selector* does not just apply to the selection of options to be nudged towards, but also to the means of nudging itself (Mele et al. 2021; Mills 2020b). For instance, perhaps—owing to the propensity of bias within a population—changing the default may be found to be a less effective means of architecting

[1] Broadly, we will ignore the rich normative debate regarding nudging and choice architecture. For some worthwhile perspectives, see Mills (2020a), Oliver (2019) and Sunstein (2014). 'Effective,' according to a strict interpretation of nudge theory, would mean that which leaves decision-makers better off, as judged by themselves. This follows from the concept of libertarian paternalism proposed by Thaler and Sunstein (2003) and further developed in their 2008 contribution. Unless otherwise clarified, we will adopt a perspective much more akin to that of *potency* emergent in the hypernudge literature (Lanzing, 2019; Yeung, 2017). Potency should generally be understood as the proportion of decision-makers who choose the option the choice architect is encouraging them to choose. Johnson and Goldstein's (2004) default option nudge for organ donation, for instance, is a highly-potent example of choice architecture because almost everyone chooses the option the choice architect wants them to choose. Choice architecture which leads to only, say, 10% of decision-makers choosing the architected option would not be very potent. Throughout, when we say an intervention or option is effective, we will generally mean potent, though we have chosen to not use the latter term because it is esoteric in comparison to most nudge literature.

choices than, say, using a social norm or a framing prompt. Prior to selecting which option should be nudged towards, therefore, choice architects may also be tasked with selecting which type of design (i.e., nudge) should be used from a set $\{X, Y, Z \ldots n\}$. This stage of selection may not always be possible or required. For instance, where a choice architect is told they can only change the default option, deliberating on alternative types of design is moot (Thaler 2021). Likewise, where technical limitations in terms of the *medium* of decision making exist (e.g., a paper form, a mobile application), such choices may also be impossible. But this caveat is largely beside the point: whether the question concerns the type of design, or the outcome, or both, choice architects engage most simply is the process of *selecting from a set of possibilities*.[2] This perspective broadly follows Mills' (2020b) framework for personalised nudging.

As with many activities, having reduced the task of architecting choices down to its most basic premise, it becomes reasonable to imagine that this task may be automated, and indeed, as examples such as Facebook's News Feed algorithm, Google's search algorithm, and Amazon's recommendation algorithm all demonstrate, *autonomous choice architects* already exist (Jameson et al. 2013; Mele et al. 2021). In this paper, we describe the autonomous choice architect as an artificial intelligence (AI). In addition to arguing that AI can take on the role as choice architect, we highlight some of the most obvious implications of this conclusion. A key challenge is that automated choice architecture arguably reduces both transparency and human accountability, and some have even gone so far as to argue that humans cannot bear full responsibility for the actions of complex machines and modern AI (Matthias 2004). While some find cause for a dissolution of responsibility when AI is used, many people not familiar with AI might intuitively hold that it is self-evident that humans remains in charge for whatever *tools* they choose to use. We agree with such a position, but argue that it is important not to rely on intuition alone when arguing in favour of human responsibility. Both policy-makers and the general public, we argue, benefit from increased knowledge of the nature of autonomous choice architects to effectively argue their case for human responsibility.

The approach we adopt in this article is largely conceptual, as our primary research interest is unpacking the functional parallels between the activities of human choice architects, and the capabilities of autonomous systems, before exploring the consequence of drawing and embracing such parallels. Our objective is to encourage reconsideration of who (or *what*) the choice architect is, as well as reflection on how the shifting form of the choice architect (from human to machine) changes the responsibilities of the choice architect. However, this article does benefit from existing alongside a growing body of critical research into the intersection of behavioural science and information technology more broadly. For instance, for a more empirical approach to similar questions raised in this article, see Mele et al. (2021). Therefore, particularly in Sect. 3, we will draw on such research and examples to evince and complement our overall conceptual assertion that the choice architect operates *as a selector*, and may be understood as viably operating autonomously, taking the form of an algorithmic or artificially intelligent system.

In Sect. 2, we offer several definitions which underpin our use of the term AI, and provide a basis for differentiating an autonomous choice architect from a traditional, *human* choice architect. In short, the position adopted in this paper is that the autonomous choice architect automates the selection and implementation of choice architecture (i.e., automatic nudging or *smart nudging*; Mele et al. 2021). In Sect. 3, we explore some examples of actual and proposed autonomous choice architects, identifying the similarities these examples share in terms of *choice architecting as a selection process*, and in terms of operating independent of direct human oversight. In Sect. 4, we consider the implications of autonomous choice architects, arguing that despite the appearance and autonomy of autonomous choice architects, this cannot and should not result in one ignoring the human actors who design, implement and control these systems. We conclude by discussing the implications of autonomous choice architects on practitioners and society.

2 Definitions

To begin our discussion, it is useful to provide a brief outline of some of the key terms which will be used throughout. We consider this exercise useful for two reasons. Firstly, terms such as 'artificial intelligence' and 'behaviour' are often used in a variety of contexts, and as such, can come to capture different concepts depending on one's perspective (Possati 2020; Turkle 2004 [1984]). For instance, in a recent review of artificial intelligence for the EU Commission, Samoili et al. (2020) identify 55 different definitions of AI. Secondly, a key aspect of our discussion is how humans and AI are similar, and thus how an AI system can come to operate as a choice architect. Thus, it is helpful to proceed with terms that establish equivalency between

[2] This perspective is something of a growing one. For instance, Benartzi (2017) highlights the growing importance of A/B testing in digital design spaces. A/B testing involves testing the effectiveness of a design choice compared to another (i.e. design A compared to B), iterating over all combinations to determine the best design from the set of possible designs. Owing to the highly-customisable nature of websites, Benartzi (2017) argues A/B testing is a vital aspect of behaviourally informed web design. Also see Weinmann, Schneider and vom Brocke (2016), Schneider, Weinmann and vom Brocke (2018), and Reinecke and Gajos (2014).

humans and machines, as well as terms which reveal important differences between humans and machines. This may produce terms some human practitioners take objection to, for instance, the notion that architecting choices is a mere selection task, rather than involving human creativity and imagination. Nevertheless, insofar as various domains now employ AI systems and algorithms to architect choices (Johnson 2021; Lavi 2017; Mele et al. 2021), it is unhelpful in understanding these entities to ignore conceptually useful points of similarity between humans and machines on the basis that they are not *perfect descriptions*.

To begin, we define behaviour, *in terms of the choice architect*, as actions in response to stimuli. Furr (2009, p. 372) has offered a more specific definition of behaviour, that being, "verbal utterances or movements that are potentially available to careful observers using normal sensory processes," yet this definition has been criticised as potentially being out-of-date given the rise of new technologies (Rauthmann 2020). For instance, specification for behaviour to be bounded by "normal sensory processes," or limited to only "verbal utterances or movements," is questionable with the emergence of fMRI machines which can be used to observe neural events, or the widespread adoption of smartphones which can track eye movements and subtle facial expressions (Valliappan et al. 2020). Our definition of behaviour is much more in keeping with perspectives from the world of AI (Silver et al. 2021; Russell 2019; Simon 1994 [1969]) which themselves have drawn from earlier theoretical models of human behaviour found in psychology (Hayek 1952; Miller 2003; Turkle 1988; von Neumann 2000 [1958]).

Where Furr's (2009) definition is more useful for our purposes is the notion of actions being "potentially available" to be observed. Following Furr (2009), just because an action could have been observed, but is not observed, does not mean that the action is not a behaviour. Insofar as the perspective we adopt throughout this paper is one of selectors choosing from sets, the notion of potential actions is desirable as it suggests an observed behaviour is just that which has been 'selected' from a set of possible actions by some criteria (Skinner 1976 [1974]). Therefore, we regard *behaviour as potentially observable actions in response to stimuli*.

We define intelligence, following Russell (2019), as *the process of selecting actions which are expected to achieve one's objectives*. Such a perspective is immediately useful in relation to our definition of behaviour. For instance, not all behaviours are necessarily intelligent. Listening to music or waving out of a window are simply actions. But if one's objective is to rehearse for a music recital, or to escape a burning building, these respective actions may be considered intelligent insofar as they are expected to achieve one's objectives (Russell 1997). Such *intelligent behaviour* may also be called rationality of some kind (Russell 1997; Schafer 2018).

The question of how an entity forms their expectations is an interesting one, and points to an important difference between humans and (existing) AI systems. de Vos (2020, p. 3) distinguishes between machines and humans based on humans' "strange capacity to reflect upon themselves," which one might describe using another word: *learning* (also see Ashby 1978; Turkle 1988). Yet the way humans learn is quite different from that of contemporary machine learning systems (Watson 2019), even if the former has inspired the latter (Silver et al. 2021; Simon 1981; von Neumann 2000 [1958]). For instance, humans do not need to see 10,000 images of a cow to correctly identify a cow, whereas AI systems currently do require such intensive training (Watson 2019) in the form of *reinforcement learning* (Silver et al. 2021). Equally, such intensive machine learning processes can produce results which exceed human results, even when a human is professionally trained in a domain. For instance, McKinney et al. (2020) report results of an AI significantly outperforming medical professionals in identifying cancerous tumours.

This leads us to our final definition, that of a machine. Marx (2013 [1867]) has offered an initial perspective on a machine, arguing it differs from a mere tool by possessing its own "motive power" (p. 257). A tool, by contrast, must be given its motive power by some other entity, usually a human. Such a perspective can also be found in more recent literature (Gunkel 2020; Turkle 2004 [1984]) and is a useful perspective to adopt for two reasons. Firstly, it provides a basis for drawing some equivalency between a human and machine—both possess their own motive power. Secondly, it invites one to ask the question, "why might one replace a human with a machine?" by providing a basis for such replacement (or automation), namely, the ability to substitute motive powers.

To answer to this question, it is because the motive power of a machine is materially different to that of a human (Marx 2013 [1867]). For instance, no human (or any number of humans) could decide which Facebook post to show a user within the time it takes for the app to load in any intelligent fashion (Gunkel 2020; Zuboff 2019). Likewise, it would likely take many weeks and several professionals to collect, analyse and output a judgement on which policy to pursue, or which medical treatment to offer, when thousands of variables need to be considered (Aonghusa and Michie 2021; McKinney et al. 2020). If one describes this material difference as *organic* (i.e., human) versus *inorganic* (i.e., machine), we define a machine as an *inorganic entity possessive of its own motive power* (Table 1).

Table 1 Summary of terminology

Term	Definition
Behaviour	*…potentially observable actions in response to stimuli*
Intelligence	*…selecting actions which are expected to achieve one's objectives*
Machine	*…an inorganic entity possessive of its own motive power*

3 AI as selection systems

In this article, we assert that, at its most basic, *descriptive* level, the role of a choice architect is to act as a selector, firstly in terms of *how* a person is influenced (i.e., which nudging strategy is used), and second in terms of *what* a person is influenced to do (i.e., what outcome/option/preference/behaviour etc. a decision-maker s nudged towards; Mills 2020b).

The *behaviour* of the choice architect (which should not be confused with the behaviour of the decision-maker) is assumed to be orientated towards an objective, and indeed, there is no example of conscious nudging which we know of whereby the choice architecture used was not part of a strategy for achieving an objective. This assumption may stand regardless of whether the nudge conforms to the typical ethical standards of nudging, whereby decision-makers are expected to be left better off by the nudge, while retaining freedom of choice (so-called *libertarian paternalism*; Thaler and Sunstein 2003). For instance, a vendor may try to nudge a decision-maker to buy a product the decision-maker will not especially enjoy, but that will be profitable for the vendor (Beggs 2016; Lavi 2017). The vendor, therefore, nudges *with an objective*, though not necessarily for the decision-maker's benefit.

Accepting this assumption, the behaviour of the choice architect can be described as intelligent, insofar as the act of architecting choices in a particular way is expected to achieve the objectives of the choice architect. Choice architects also learn, or are expected to learn, from previous experiments with changing choice architecture. Some recent reviews include Della Vigna and Linos (2020), who review the effectiveness of nudge randomised-controlled trials (RCTs) across two so-called 'nudge units;' Beshears and Kosowsky (2020) who review 174 nudge studies to evaluate the average effect size of different nudge strategies; and Jachimowicz et al. (2019), who review 58 studies specifically investigating the default option nudge to determine the effect size associated with this specific nudge. There have also been recent calls to consider experimental practices in choice-architectural design (John 2021), to consider strategies for scaling nudge interventions (Al-Ubaydli et al. 2021), and for widespread adoption of A/B testing methods (Benartzi 2017).

Human choice architects are also autonomous insofar as they have their own motive power. Yet, as the history of the computer and data collection attests, the material difference in motive power between humans and machines means that humans may be displaced by machines (Zuboff 1988). Indeed, one key consideration for the automation of previously human processes is the sophistication of the task (Brynjolfsson and McAfee 2014). Tasks which can be routinised and described in terms of computational (i.e., mathematical) logic are often prime candidates for automation, and as Simon (1994 [1969]) has argued, the abstraction of a task into a series of instructions often leads to more efficiency (e.g., faster results, more accurate results). The description of the choice architect as a selector is purposely used because it is a description which lends itself to a discussion of automation, while the arguments provided above regarding the efficiency of automated, often algorithmic processes, is also increasingly drawn upon in the discussion of architecting choices (Aonghusa and Michie 2021; Mele et al. 2021; Yeung 2017).

Consider the following example. One common objective which choice architects hold is achieving greater efficiency from nudging, usually understood to mean more people achieving outcomes which leave them, "better off, as judged by themselves" (Thaler and Sunstein 2008, p. 8). Owing to the computational limitations of human choice architects, most nudges must adopt a 'one-size-fits-all' approach, which may be effective on aggregate but may not be *optimal* (Mills 2020b; Sunstein 2012). A strategy which is closer to the 'optimal,' as is increasingly being argued, would be to personalise choice architecture so that different people are nudged in ways which respect their individual differences (Lanzing 2019; Mills, 2020b; Peer et al. 2020; Porat and Strahilevitz 2014; Sunstein 2012, 2013; Yeung 2017).

At its core, this is a problem of selecting choice architectural designs dependent on the person being nudged (i.e., an input variable), and a problem which would seem resolvable by the introduction of machines (Sunstein 2013; Yeung 2017). Sunstein (2013) argues that the central technical challenge for personalised nudging is utilising *heterogeneity data*, or data which is required to determine differences between individuals (Mills 2020b). Additionally, Yeung (2017) argues that automated systems such as AI are valuable because they can respond much more rapidly to feedback than any human choice architect could; an argument which, essentially, is also a data analysis problem which emerges from sustained (i.e., intertemporal) data collection. Aonghusa and Michie (2021) have also argued that the

amount of data which can now be drawn upon, and which is necessary to achieve precise and personalised choice architecture, is beyond reasonable human capacities to analyse. Therefore, they suggest AI systems may be better suited for data analysis within behavioural science. In short, autonomous choice architects could learn to be better selectors of choice architecture than human choice architects, could do this faster, and could resolve difficult data challenges which are currently beyond human capacities, with personalised nudging being the prime example of these advantages in combination.

To further illustrate how choice architecture may be arranged by an autonomous, AI system, we will briefly discuss how AI systems can and are acting as selectors of choice architecture in the areas of (a) policy evaluation; (b) ecommerce, and (c) information environments. We have chosen these areas for their historic prominence within the behavioural science literature (Sanders et al. 2018) and for their continued relevance within the contemporary digital economy. This discussion is given to be illustrative, not definitive or to develop any wider framework. For more complete frameworks regarding the interaction of behavioural science and technology for nudging, see for instance Mele et al. (2021), Schneider et al. (2018) and Villanova et al. (2021).

3.1 Policymaking

Aonghusa and Michie (2021) present a developed academic contribution to the question of AI and behavioural science, building from a theoretical proposal first offered in 2017 (Michie et al. 2017). Concerned with how behavioural science can be used to achieve better public health policy, while aware of a vast corpus of behavioural health literature which exists, Aonghua and Michie (2021) present the results of a collaborative effort to design an AI system which uses this vast corpus to make predictions about the effectiveness of a particular intervention given various criteria, such as the age of the target population and the setting in which the intervention is to be used. Having specified these details, they report a computational system capable of producing statements such as:

> "For a population with **Minimum Age 68** and **Maximum Age 79, Goal-Setting Behaviour Change Technique** [*sic*] in a **Care Home Facility** setting is likely to lead to **5%** of subjects stopping smoking for at least 3 months." (p. 944, original emphasis)

The choice architect does not necessarily know *why* a given intervention with a given set of variables is predicted to produce, in this example, a 5% effect, but the choice architect is empowered in a way which, prior to the AI, may not have been possible. By contrast, consider a human choice architect engaging in this activity. Firstly, they would have to identify the relevant information from a vast body of literature,[3] secondly, determine which information is generally relevant, and which information is *specifically* relevant to the scenario being considered, and finally use that information in some way to reliably predict the effectiveness of the intervention. This is not to downplay the likely significant development-time for such an AI system, but the AI system Aonghusa and Michie (2021) discuss, once developed, can evaluate choice architecture much more rapidly than could be achieved by a human, and serves as an intriguing example of an autonomous choice architect.

Of course, as noted above, the system described by Aonghusa and Michie (2021) is not autonomous in the sense that it can *implement* the choice architecture. At present, in terms of implementing choice architecture in a policy space, the AI is much more a *tool* for human choice architects, than a truly *autonomous choice architect*. Nevertheless, as experiments such as that by Peer et al. (2020) demonstrate, it seems reasonable to expect that in the coming years such policy evaluation may come to be embedded seamlessly as part of a personalised, choice architectural experience, and at this point, the system discussed by Aonghusa and Michie (2021) may qualify as *wholly autonomous*.

3.2 eCommerce

Mills et al. (2021) present a model of choice architecture and spending behaviour in their concept of *SpendTech*. They define SpendTech as, "a range of technologies used in conjunction with behavioural insights to induce desired spending behaviours at a decision-point" (p. 3). The authors argue that datafied consumers can be modelled using AI and machine learning technologies as probabilistic subjects, with variables such as the information presented, the biases of the consumer, the individual differences of the consumer (see *heterogeneity*; Sunstein, 2012), and the product differences across purchases, all constituting inputs which can be evaluated by an AI system tasked with optimising the likelihood of a consumer purchasing a product through the architecting of choice environments. Some examples given by the authors include:

1. Recommendation algorithms learning from a user's previous purchases, as well as purchases of people similar to the user, to select from a set of possible products that product which is predicted to have the highest chance of being bought by a particular user, and then recommending it.

[3] Aonghusa and Michie (2021) do not report the size of the corpus.

2. AI analysing purchasing habits over time to determine times within, say, a monthly cycle when a person is more inclined to splurge or more inclined to be frugal, and adjust the use of choice architectural techniques such as dynamic pricing and fear-of-missing-out to maximise effectiveness.
3. Through Big Data and constant data analysis, automated systems identifying trends and patterns between products, purchasers, and a litany of other factors to draw inferences about customers, for instance by predicting customers who are expectant mothers and configuring their choice environments to feature pregnancy-orientated products.

'SpendTech' is but one perspective on the use of AI systems to select optimal choice architecture within the ecommerce domain. Villanova et al. (2021) discuss the notion of 'just-in-time' messaging, retail messages which are tailored in language and timing to be maximally persuasive to a potential customer, while Matz et al. (2017) provide empirical evidence of the effectiveness of online advertisement targeting using personality predictions, which is elaborated on by Matz and Netzer (2017). Smith et al. (2020) offer the notion of *exogeneous cognition*, where rather than decision-maker's actively thinking about their spending decisions (e.g., endogenous cognition), choice architecture curated by automated technical systems remove the 'burden' of thought and ease the spending decision (Frischmann and Selinger, 2018; Mele et al. 2021, 2019). Hauser et al. (2009, p. 202) discuss the notion of *website morphing*, which "involves automatically matching the basic "look and feel" of a website, not just the content, to cognitive styles" with the aim of increasing customer engagement and spending.

Often, these systems can, or *do*, function as entirely autonomous choice architects. It is not simply that the AI system suggests, say, what product to put in a recommendation list, before a human choice architect *actually implements the suggestion*. Instead, the AI implements its selection itself. For instance, Amazon and other vendors occasional receive mockery and concern when AI-generated products go viral (Wiggers, 2020). These products, typically those that cost relatively little to make highly-customised, such as T-shirts and coffee mugs, are generated automatically based on the notion that a very specific product is likely to be highly desirable to very few people, but profitable insofar as the product can be targeted reliably at *only these people*. Despite this difference, there are clear functional similarities to be found, such as both examples demonstrating the need to analyse vast amounts of data which are far beyond human capacities, but well-suited to AI systems.

3.3 Information environments

Sharot and Sunstein (2020) explore how people use information which is given to them, and how much information *should* be given to people to influence their behaviour. They argue there are three types of utility which should be considered when determining if a piece of information should be shown or not: *instrumental utility* ("Action")[4]; *hedonic utility* ("Affect")[5]; and *cognitive utility* ("Cognitive").[6] Based on these inputs, which they argue are subject to individual differences, the subjective value of any given piece of information can be determined as a prediction: "These estimates [of utility] are integrated into a computation of the value of information" (p. 16). Ultimately, this predicted value of information can be used, very basically to determine if a piece of information should be shown (i.e., is it expected to increase or decrease utility?), or more dynamically, to determine the *relative* value of information compared to other information. In short, Sharot and Sunstein's (2020) concept follows a similar pattern seen previously: the three utility estimates, along with various measures of individual difference, are input variables subjected to some undefined, "computation," to determine an estimate of the value of that particular piece of information, which can then be used to select choice architecture.

Similar notions have been proposed by Thaler and Tucker (2013), whose subsequently prescient concept of the "choice engine" (p. 44) describes the use of data and algorithms to determine, on an individual-level, which information should be given in various disclosure documents to maximise individual understanding.[7] Emerging concepts in the financial world, such as *Robo-Advice*, follow a similar principle of providing financial advice based on an individual's financial circumstances and preferences. Finally, and perhaps most obviously, the Facebook News Feed algorithm, Google's search algorithm, and the YouTube recommendation algorithm may all be described in terms of using input data to curate the information shown to users to maximise an outcome. Crucial to this discussion, regardless of the example, is that choice architecture in terms of information disclosure can be seen as a *selection* process, with many instances where a vast amount of information *could be provided*, but a smaller, more-tailored amount of information is predicted

[4] "Will the knowledge help, hinder or have no influence on my ability to make decisions to increase reward and avoid harm?".

[5] "Will the information induce positive or negative feelings, or will it have no influence on my affect?".

[6] "Will information improve my ability to comprehend and anticipate reality?".

[7] Also see Johnson (2021) for a more recent, but complimentary, use of the term 'choice engine.'

to be closer to optimal (Sharot and Sunstein 2020; Thaler and Tucker 2013).

Insofar as the examples given, and indeed insofar as Sharot and Sunstein (2020) explain their concept given the subjectivity of information's value and estimations of utility, the choice architecture of information disclosure lends itself towards the autonomous choice architect. A platform like Facebook, for instance, does not have a human choice architect determining which content each user should see; an autonomous choice architect evaluates all possible options, selects those options expected to maximise an objective (e.g., click-through-rate; retention), and implements this selection.

4 Accountability and autonomous choice architects

Our discussion has frequently drawn on contrasts between human choice architects and autonomous choice architects, but often only insofar as we argue that machines can replicate and surpass the functions of human choice architects. However, thus far a discussion of accountability has been missing. For instance, who is responsible for an algorithm which automatically architects choices to influence behaviour? As discussion of technology-linked nudging, such as hypernudging (Yeung 2017), Big Data nudging (Sætra 2019), digital nudging (Weinmann et al. 2016), smart nudging (Mele et al. 2021), and generally AI nudging, has developed, so too has the risk that *human* responsibility for the actions of these autonomous systems will slip out of focus. Some might intuitively see it as self-evident that using an AI system as a choice architect is akin to using any other tool, and that it entails no dissolution of human accountability. However, increased machine complexity and increasing machine autonomy has given rise to arguments implying that we must take seriously the possibility that machines that act intelligently can also be held accountable for the actions they take. We agree with those that intuitively hold that humans must be held accountable for the actions of machines, but merely assuming that this is self-evident leaves openings for opponents of such a view. To effectively argue in favour of human accountability in a world of autonomous choice architects, the preceding considerations about what AI is and is not capable of, and an understanding of how to bring this into an evaluation of potential shifts in accountability, is required.

As AI systems are used to either assist or replace human decision-makers, some argue that a "responsibility-gap" is thus created (Matthias 2004). This gap, it is argued, emerges because it is wrong, unfair, or otherwise problematic to attribute responsibility for machine actions to humans who cannot "anticipate, completely control, or answer for" these actions (Gunkel 2020). The problem is associated with the complexity and the autonomy of the machines. As machines become more complex, at some point complexity surpasses a human's ability to understand the exact workings of the machine (Rahwan et al. 2019; Pedersen and Johansen 2020). In addition to complexity, certain systems use error as a method for learning and reconfiguration, and thus become inherently unpredictable (Matthias 2004). This is particularly relevant for unsupervised AI systems, where AI is tasked with identifying patterns and connections unknown to *anyone*, but remains an issue for supervised AI also, where understanding the process of learning can remain opaque (Pedersen and Johansen 2020). This produces the classic 'black box' problem of AI, where questions such as "what is the AI doing?" and "why is it doing that?" become extremely hard to answer (Durrell 2016; Pasquale 2015; Turkle 1988). Insofar as it is relevant to this discussion, the question, then, is: do complexity and unpredictability relieve humans of responsibility for entities such as autonomous choice architects?

We approach this question by considering how nudging can be automated and connected to AI systems. First of all, we argue that nudging is hardly ever *direct* in the sense that the choice architect directly interacts with the decision-maker. In changing the default on a form, or installing a poster in a cafeteria, the choice architect's actions are expressed via the form, or via the poster. Thus, the choice architect *is often directly involved in creating and modifying* the choice architecture involved, but *not* in directly interacting with the decision-maker. From this perspective, the apparent *indirectness* arising from the autonomous choice architect is irrelevant to the question of responsibility.

Instead, a "veil of complexity" (Sætra 2021a) obscures the involvement of humans in automated and algorithmic nudging. In the first instance, AI systems constitute an additional layer of indirectness, as human choice architects are not directly involved in choice architecture and are not directly involved in applying it in settings in which the targets of nudges encounter them. Such 'distance' may be compounded when multiple choice architects are involved in designing the AI, making it harder to confidently determine who is responsible for what (Matthias 2004). In the second instance, a human choice architect may not be able to understand or explain the actions of the autonomous choice architect, and in a hypothetical confrontation between themselves and a decision-maker, may feel inclined to declare ignorance of the action and shirk responsibility. Yet, in lifting this veil of complexity, we argue that little actually changes in terms of *who is responsible*.

Classical examples of choice architecture manipulation accepted as nudges are the positioning of items in a store (near the check-out counter, for example; Thaler and Sunstein 2008), the positioning of items of restaurant menus

(Dayan and Bar-Hillel 2011), and changing default options to encourage organ donation (Johnson and Goldstein 2004). All of these nudges can be performed manually, by a choice architect who selects how items are placed in the store, designs the menu, or changes the default option from opt-in to opt-out. This basic, or *fully manual*, form of nudging, is static and requires another manual action by a choice architect to change the nudge. Furthermore, all targets of the nudge encounter it in the same place, time, and manner if they behave similarly (Mills 2020b).

To both reduce the workload for the choice architect and to make the nudges more effective, choice architecture can be *automated*. In the case of the restaurant menu, for example, the restaurant could use an app and digital menus in-store to present different menus at different times of the day. This could be based on the choice architect knowing that daytime customers require a different sort of nudge than evening customers to maximize profits (or to avoid unhealthy eating—the goal is not of significance for the principle involved). From 10 a.m. until 5 p.m., one menu is shown, and this switches automatically to a different menu from 5 p.m. until closing. We argue that, while automated, there are no reasons to argue that the choice architect is now less responsible for the nudges or their effects than they would have been had the nudge been fully manual.

What, then, if the restaurant focused exclusively on app-based and individually delivered menus, and implemented a system for registering users and analysing information on these users with the use of intelligent systems? This could be done to maximally harness the potential of personalised choice architecture (Mills 2020b). For the sake of the argument, we assume that the restaurant is operated by a company that also runs many different types of stores and services, and that they already have a vast array of data points of each customer. Rather than manually deciding what sort of menu to show at a given time, or manually tailoring the menu to each person, the company decides to use unsupervised deep learning algorithms to decide how menus are presented. The ultimate *goal* is provided, to maximize long-term profits in this case, and the AI system provides the personalised menus in-app. No one, not the customer, and no one in the company or those developing the app or algorithm, (a) understands or can explain why individual X gets menu Y, and (b) no one has been directly involved in the particular choice architecture presented to the customers. This, one might argue, is how the autonomous choice architect is born.

The latter form of system involves distancing the human choice architect from the actual end results, but we argue that despite the unpredictability and opaqueness of such systems, it is imperative to recognise humans' role in the process as choice architects. While sophisticated deep learning systems introduce a veil of complexity, a proper understanding of these systems, both of the algorithms and the data, allow us to remove this veil. We emphasise five keyways in which human influence remains clear. First, humans program the systems and algorithms involved, and this will always entail making a range of choices regarding how the systems will end up making decisions. Secondly, humans decide how, where, and when these systems are to be applied, such as in a particular restaurant. Thirdly, humans directly instruct these systems to optimise based on selected variables. Fourthly, humans are involved in deciding which variables or factors, which parts of the choice architecture, are manipulable by the AI system. Fifthly, and finally, humans are involved in a wide range of actions that shape and influence the generation, selection, and codification of the data used by these systems (Sætra 2018).

Humans are irrevocably involved, even when autonomous choice architects are used. Granted, the involvement is several steps removed, and *appears* more indirect, and may be orientated more to the *design of a machine* than the *design of choice environments*, but insofar as responsibility is a function of control (Gunkel 2020; Matthias 2004), and insofar as control over an autonomous choice architect can be attributed to whoever sets the *parameters* of it, *humans* remain responsible (Sætra 2021a). If machines have become autonomous choice architects, this is, if anything, delegated or automated choice architecture, and human beings remain responsible.

Another interesting question to consider is *why does this matter for nudging*? In terms of AI, an autonomous choice architect is not really unique, and in terms of responsibility, our discussion of responsibility could apply to an AI which is nudging, or to an AI which is driving a car or conducting heart surgery. Yet, one might make a case for special consideration in regard to using AI to design choice architecture.

Nudges are supposed to allow individuals to "go their own way" (Thaler and Sunstein 2008, p. 5). They are, in a sense, *suggestions* which a decision-maker should be able to ignore (Madrian and Shea 2001), even if they are suggestions designed to exploit human bias. Much objection has already been made towards what we have here called manual nudging (Rebonato 2014; Mitchell 2005), and we do not wish to restate these arguments, particularly in regard to freedom of choice. Yet, the very nature of autonomous choice architects, owing to their mechanical motive power, mean they are able to respond faster to decision-maker feedback, able to integrate a more perfect image of the decision-maker in the design selection, and inevitably able to attune their selections to a specific

decision-maker in *real-time* (Sætra 2019; Yeung 2017). Automatic nudging, therefore, may only *nominally* accord to nudging insofar as it allows one to "go their own way," as automatic nudging can design choice architecture, collect feedback, and change choice architecture so quickly, and so efficiently, that a decision-maker may never be able to escape its influence.[8] Insofar as such systems continue to be described as 'nudges,' while not *substantially* allowing people to 'go their own way,' autonomous choice architecture may present a significant ethical challenge for the wider nudging programme.[9]

The consequence of this is it compounds the veil of complexity by allowing a choice architect to excuse their influence on the basis that decision-makers can always ignore them. If a decision-maker is dissatisfied, so the argument might go, *it is their own fault for following the nudge*. Of course, if one's ability to *not follow* is compromised, such an excuse is not valid, but could still be made in an attempt to avoid responsibility.

Finally, a long-standing criticism of what we have called *manual nudging* is that there is no particular reason to believe that a choice architect should know any better or any worse than a decision-maker regarding what the 'best' or 'correct' choice is (Selinger and Whyte 2010). This is largely because a) people are often different, with different preferences and motivations, and b) the choice architect is often removed from those they are architecting choices for, and may even be biased themselves (Rebonato 2014). In some ways, autonomous choice architects may function as a response to this otherwise tricky criticism of nudging by utilising individual-level data to personalise choice environments (Mills 2020b; Porat and Strahilevitz 2014; Yeung 2017; Mele et al. 2021).

However, autonomous choice architects may *compound* the problems of knowledge, competence and trust discussed by Selinger and Whyte (2010) by robbing human choice architects of any semblance of the possibility of explaining to a decision-maker why they were nudged in the manner that they were. Insofar as we argue human choice architects are ultimately responsible and accountable for the actions of autonomous choice architects, autonomous choice architects may, rather than empowering humans, act as an impediment to effective discourse between choice architect and decision-maker about why and how nudges should be used.

[8] Mills (2021) describes such a situation not as a *nudge*, but—borrowing from Yeung (2017)—a *hypernudging system*: many nudges, connected via an algorithm, in such a way that a decision-maker may reject a specific nudge, but will always *be nudged somehow*.

[9] A tentative term for such a phenomenon may be *nudgewashing* – coercion, justified under the guise of being able to 'go your own way', but designed in such a way as to undermine one's freedom to not be coerced.

5 Implications for practitioners and society

Autonomous choice architects already exist, and insofar as they allow many services to be automated and to convey social benefit via, say, tailored information, disclosures, or default options, may also be described perhaps more positively than we have done in this article (see Mele et al. 2021). Yet many of these benefits-speed of reconfiguration, ability to integrate individual-level data, responsiveness to feedback in real-time-produce implications which prompt reconsideration of the implications of the role of the choice architect.

From the practitioner's perspective, the autonomous choice architect may be seen as a threat, not so much from the perspective of *automating away their job*-research into behavioural science and nudge design seems likely to continue to be a human-dominated domain for the moment-but from the shifting responsibility they bring. Practitioners need to understand the systems they are implementing, not only on a technical level (insofar as this is possible; Burrell 2016), but also in terms of their responsibility *over what the system does* and *why*. We can break these questions down further into more specific inquiries:

1. Given a random decision-maker, could I predict and explain how they will be nudged?
2. If this decision-maker wanted to 'go their own way,' would they meaningfully be able to do so?
3. What decisions have I made about the decision-maker, and what consultation have I had with the decision-maker, and what recourse do they have?
4. Do I consider the actions of an autonomous choice architect which I implement to be my responsibility?

These questions likely vary from context to context (e.g., a private-sector choice architect may have different obligations to stakeholders compared to a public-sector choice architect; Beggs 2016), but we consider them a worthwhile initial set on questions for the behavioural (and data) science community to begin with.

In terms of implications for society more broadly, recent developments show that regulators are adopting a stance that might drastically limit uncritical choice architecture automation. The European Union's GDPR framework already introduced a right to explanation, or at least a right to meaningful information about the logic involved in automated decision making (Selbst and Powles 2018). More recently, however, the European Commission has proposed a new AI regulation that goes further in limiting how AI can be used for changing people's behaviour (European Commission 2021). Here they propose a ban of AI systems thought to entail "unacceptable risk", and one of the examples mentioned is

"systems or applications that manipulate human behaviour to circumvent users' free will". Whether or not autonomous choice architects *manipulate* or not is a contested question (Sætra 2021b), but the phrasing in the proposal is vague enough for the systems we discuss in this article to potentially be included. In addition, the proposal suggests that certain contexts are "high-risk", such as education, critical infrastructure, and law enforcement. They proceed to suggest that any use of AI in such contexts should be associated with strict requirements related to, for example, "appropriate human oversight", "adequate risk assessment and mitigation systems", "clear and adequate information to the user", "high quality of the datasets," and a "high level of robustness, security and accuracy" (European Commission 2021). Regardless of the outcome, it seems clear that governments are increasingly taking an interest in mitigating potentially negative consequences of AI. This is both a consequence of increasing awareness of the implications associated with particular technologies, such as the ones discussed in this article, but also to a potential shift in the willingness to allow the political domain to limit the relatively free application of technological innovations (Sætra and Fosch-Villaronga, 2021).

References

Al-Ubaydli O, Lee MS, List JA, Mackevicius CL, Suskind D (2021) How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling. Behav Public Policy 5(1):2–49

Aonghusa PM, Michie S (2021) Artificial intelligence and behavioral science through the looking glass: challenges for real-world application. Ann Behav Med 54:942–947

Ashby WR (1978) Design for a brain. Chapman and Hall

Beggs J (2016) Private-sector nudging: the good, the bad, and the uncertain. In: Abdukadirov S (ed) Nudge theory in action (2016). Palgrave Macmillan, London

Benartzi S (2017) The smarter screen: surprising ways to influence and improve online behavior. Portfolio Books

Beshears J, Kosowsky H (2020) Nudging: progress to date and future directions. Organ Behav Hum Decis Process 161:3–19

Brynjolfsson E, McAfee A (2014) The second machine age: work, progress, and prosperity in a time of brilliant technologies. W. W. Norton and Company

Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. https://doi.org/10.1177/2053951715622512

Dayan E, Bar-Hillel M (2011) Nudge to nobesity II: Menu positions influence food orders. Judgm Decis Mak 6(4):333–342

de Vos J (2020) The digitalisation of (inter) subjectivity. Routledge

Della Vigna S, Linos E (2020) RCTs to scale: comprehensive evidence from two nudge units. https://eml.berkeley.edu/~sdellavi/wp/NudgeToScale2020-05-09.pdf. (**Date accessed: 24/03/2021**)

European Commission (2021) Europe fit for the digital age: commission proposes new rules and actions for excellence and trust in artificial intelligence. https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682. (**Date accessed: 22/10/2021**)

Frischmann B, Selinger E (2018) Re-engineering humanity. Cambridge University Press

Furr MR (2009) Personality psychology as a truly behavioural science. Eur J Pers 23:369–401

Gunkel DJ (2020) Mind the gap: responsible robotics and the problem of responsibility. Ethics Inf Technol 22(4):307–320

Hauser JR, Urban GL, Liberali G, Braun M (2009) Website morphing. Mark Sci 28(2):202–223

Hausman DM, Welch B (2010) Debate: to nudge or not to Nudge. J Polit Philos 18(1):123–136

Hayek FA (1952) The sensory order: an inquiry into the foundations of theoretical psychology. Chicago University Press

Helbing D (2015) Societal, economics, ethical and legal challenges of the digital revolution: from big data to deep learning, artificial intelligence, and manipulative technologies. SSRN. https://ssrn.com/abstract=2594352. (**Date accessed: 21/10/2021**)

Jachimowicz JM, Duncan S, Weber EU, Johnson EJ (2019) When and why defaults influence decisions: a meta-analysis of default effects. Behav Public Policy 3(2):159–186

Jameson A, Berendt B, Gabrielli S, Cena F, Gena C, Vernero F, Reinecke K (2013) Choice architecture for human-computer interaction. Found Trends Hum-Comput Interact 7(1–2):1–235

John P (2021) Let's walk before we can run: the uncertain demand from policymakers for trials. Behav Public Policy 5(1):112–116

Johnson EJ, Goldstein DG (2004) Defaults and donation decisions. Transplantation 78:1713–1716

Johnson EJ, Shu SB, Dellaert BGC, Fox C, Goldstein DG, Häubl G, Larrick RP, Payne JW, Peters E, Schkade D, Wansink B, Weber EU (2012) Beyond nudges: tools of a choice architecture. Mark Lett 23:487–504

Johnson EJ (2021) How Netflix's choice engine drives its business. Behavioral Scientist. https://behavioralscientist.org/how-the-netflix-choice-engine-tries-to-maximize-happiness-per-dollar-spent_ux_ui/. (**Date accessed: 22/10/2021**)

Lanzing M (2019) "Strongly Recommended" revisiting decisional privacy to judge hypernudging in self-tracking technologies. Philos Technol 32:549–568

Lavi M (2017) Evil Nudges. J Entertain Technol Law 21(1):1–93

Luckerson V (2015) Here's how facebook's news feed actually works. Time Magazine. https://time.com/collection-post/3950525/facebook-news-feed-algorithm/. (**Date accessed: 08/03/2021**)

Madrain BC, Shea DF (2001) The power of suggestion: inertia in 401(k) participation and savings behavior. Q J Econ 116(4):1149–1187

Marx K (2013 [1867]) Capital. Wordsworth

Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6(3):175–183

Matz SC, Netzer O (2017) Using big data as a window into consumers' psychology. Curr Opin Behav Sci 18:7–12

Matz SC, Kosinski M, Nave G, Stillwell DJ (2017) Psychological targeting as an effective approach to digital mass persuasion. PNAS 114(48):12714–12719

McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, Etemadi M, Garcia-Vicente F, Gilbert FJ, Halling-Brown M, Hassabis D, Jansen S, Karthikesalingam A, Kelly CJ, King D, Ledsam JR, Melnick D, Mostofi H, Peng L, Reicher JJ, Romera-Paredes B, Sidebottom R, Suleyman M, Tse D, Young KC, de Fauw J, Shetty S (2020) International evaluation of an AI system for breast cancer screening. Nature 577:89–94

Mele C, Polese F, Gummesson E (2019) Once upon a time… technology: a fairy tale or a marketing story? J Mark Manag 35(11–12):965–973

Mele C, Spena TR, Kaartemo V, Marzullo ML (2021) Smart nudging: How cognitive technologies enable choice architecture for value co-creation. J Bus Res 129:949–960

Michie S, Thomas J, Johnston M, Aonghusa PM, Shawe-Taylor J, Kelly MP, Deleris LA, Finnerty AN, Marques MM, Norris E, O'Mara-Eves A, West R (2017) The Human Behaviour-Change Project: Harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation. Implement Sci 12(121):1–12

Miller GA (2003) The cognitive revolution: a historical perspective. Trends Cogn Sci 7(3):141–144

Mills S (2020a) Nudge/sludge symmetry: on the relationship between nudge and sludge and the resulting ontological, normative and transparency implications. Behav Public Policy. https://doi.org/10.1017/bpp.2020.61

Mills S (2020b) Personalized Nudging. Behav Public Policy. https://doi.org/10.1017/bpp.2020.7

Mills S, Whittle R, Brown G (2021) SpendTech. Unpublished Manuscript.

Mills S (2021) Into hyperspace: a critique of hypernudge' SSRN. https://papers.ssrn.com/abstract=3802614. Accessed 03 Nov 2021

Mitchell G (2005) Libertarian paternalism is an oxymoron. Northwest Univ Law Rev 99(3):1245–1278

von Neumann J (2000 [1958]) The computer and the brain. Yale University Press

Oliver A (2019) Towards a new political economy of behavioral public policy. Public Adm Rev 79(6):917–924

Pasquale F (2015) The black box society: the secret algorithms that control money and information. Harvard University Press

Pedersen T, Johansen C (2020) Behavioural artificial intelligence: an agenda for systematic empirical studies of artificial inference. AI Soc 35:519–532

Peer E, Egelman S, Harbach M, Malkin N, Mathur A, Frik A (2020) Nudge me right: personalizing online security nudges to people's decision-making styles. Comput Hum Behav 109:e.106347

Porat A, Strahilevitz LJ (2014) Personalizing default rules and disclosure with big data. Mich Law Rev 112(8):1417–1478

Possati LM (2020) Algorithmic unconscious: why psychoanalysis helps in understanding AI. Palgrave Commun. https://doi.org/10.1057/s41599-020-0445-0

Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon J, Breazeal C, Crandall JW, Christakis NA, Couzin ID, Jackson MO, Jennings NR, Kamar E, Kloumann IM, Larochelle H, Lazer D, McElreath R, Mislove A, Parkes DC, Pentland A, Roberts ME, Shariff A, Tenenbaum JB, Wellman M (2019) Machine behaviour. Nature 568:477–486

Rauthmann JF (2020) A (More) Behavioural science of personality in the AE of multi-modal sensing, big data, machine learning, and artificial intelligence. Eur J Pers 34:593–598

Rebonato R (2014) A Critical Assessment of Libertarian Paternalism. J Consum Policy 37:357–396

Reinecke K, Gajos KZ (2014) Quantifying visual preferences around the world. In: CHI '14: Proceedings of the SIGCHI conference on human factors in computing systems https://doi.org/10.1145/2556288.2557052

Russell SJ (1997) Rationality and intelligence. Artif Intell 94:57–77

Russell SJ (2019) Human compatible: AI and the problem of control. Penguin Books

Sætra HS (2018) Science as a vocation in the era of big data: the philosophy of science behind big data and humanity's continued part in science. Integr Psychol Behav Sci 52(4):508–522

Sætra HS (2019) When nudge comes to shove: liberty and nudging in the era of big data. Technol Soc 63:e.101130

Sætra HS (2021a) Confounding complexity of machine action: a Hobbesian account of machine responsibility. Int J Technoethics (IJT) 12(1):87–100

Sætra HS (2021b) Big Data's threat to liberty: surveillance, nudging, and the curation of information. Elsevier

Sætra HS, Fosch-Villaronga E (2021) Research in AI has implications for society: How do we respond? Morals Mach 1(1):60–73

Samoili S, López CM, Gómez E, de Prato G, Martínez-Plumed F, Delipetrev B (2021) AI Watch: Defining Artificial Intelligence Towards an Operational Definition and Taxonomy of Artificial Intelligence. JRC Technical Reports EUR 30117 EN. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC118163/jrc118163_ai_watch._defining_artificial_intelligence_1.pdf. (**Date accessed: 25/02/2021**)

Sanders M, Snijders V, Hallsworth M (2018) Behavioural science and policy: where are we now and where are we going? Behav Public Policy 2(2):144–167

Schafer K (2018) A brief history of rationality: reason, reasonableness, rationality, and reasons. Manuscrito 41(4):501–529

Schneider C, Weinmann M, vom Brocke J (2018) Digital nudging: guiding choices by using interface design. Commun ACM 61(7):67–73

Selbst A, Powles J (2018) "Meaningful Information" and the right to explanation. In: conference on fairness, accountability and transparency (pp 48–48). PMLR.

Selinger E, Whyte KP (2010) Competence and trust in choice architecture. Knowl Technol Policy 23:461–482

Sharot T, Sunstein C (2020) How people decide what they want to know. Nat Hum Behav 4:14–19

Silver D, Singh S, Precup D, Sutton RS (2021) Reward is enough. Artif Intell 299:e.103535

Simon HA (1955) A behavioral model of rational choice. Q J Econ 69(1):99–118

Simon HA (1981) Information-processing models of cognition. J Am Soc Inf Sci 32(5):364–377

Simon HA (1994) [1969] 'The sciences of the artificial,' 2nd edn. MIT University Press

Skinner BF (1976 [1974]) About Behaviorism. Vintage

Smith A, Harvey J, Goulding J, Smith G, Sparks L (2020) Exogenous cognition and cognitive state theory: the plexus of consumer analytics and decision-making. Mark Theory. https://doi.org/10.1177/1470593120964947

Sunstein C (2013) The storrs lectures: behavioral economics and paternalism. Yale Law J 122:1826–1899

Sunstein C (2014) Why Nudge? The politics of libertarian paternalism. Yale University Press

Sunstein C (2012) Impersonal default rules vs. active choices vs. personalized default rules: a triptych. SSRN. https://ssrn.com/abstract=2171343. (**Date accessed: 24/03/2021**)

Sunstein C (2017) Misconceptions about Nudges. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3033101. (**Date accessed: 15/04/2020**)

Sunstein C (2018) Algorithms, correcting biases. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3300171. (**Date accessed: 24/03/2021**)
Susser D, Roessler B, Nissenbaum H (2019) Online manipulation: hidden influences in a digital world. Georgetown Law Technol Rev 4(1):1–45
Thaler R (2021) What's next for nudging and choice architecture? Organ Behav Hum Decis Process 163:4–5
Thaler R, Sunstein C (2003) Libertarian paternalism. Am Econ Rev 93(2):175–179
Thaler R, Sunstein C (2008) Nudge: improving decisions about health, wealth and happiness. Penguin Books
Thaler R, Tucker W (2013) Smarter information, smarter consumers. Harvard Bus Rev 91(1–2):44–54
Thaler R, Sunstein C, Balz J (2012) Choice architecture. In: Shafir E (ed) The behavioral foundations of public policy (2012). Princeton University Press
Turkle S (1988) Artificial intelligence and psychoanalysis: a new alliance. Dædalus 117(1):241–268
Turkle S (2004 [1984]) The second self: computers and the human spirit, 20th anniversary. MIT Press
Villanova D, Bodapati AV, Puccinelli NM, Tsiros M, Goodstein RC, Kushwaha T, Suri R, Ho H, Brandon R, Hatfield C (2021) Retailer marketing communications in the digital age: getting the right message to the right shopper at the right time. J Retail 97(1):116–132
Villiappan N, Dai N, Steinberg E, He J, Rogers K, Ramachandran V, Xu P, Shojaeizadeh M, Guo L, Kohlhoff K, Navalpakkam V (2020) Accelerating eye movement research via accurate and affordable smartphone eye tracking. Nat Commun 11(4553):1–12
Vincent J (2020) Facebook is now using AI to sort content for quicker moderation. The Verge. https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation. (**Date accessed: 22/03/2021**)
Watson D (2019) The rhetoric and reality of anthropomorphism in artificial intelligence. Mind Mach 29:417–440
Weinmann M, Schneider C, vom Brocke J (2016) Digital nudging. SSRN. ssrn.com/abstract-2708250. (**Date accessed: 30/01/2020**)
Wiener N (2013 [1948]) Cybernetics or, control and communication in the animal and the machine. Martino Publishing
Wiggers K (2020) Amazon's AI generates images of clothing to match text queries. Venture Beat. https://venturebeat.com/2020/03/02/amazons-ai-generates-images-of-clothing-to-match-text-queries/. (**Date accessed: 24/03/2021**)
Yeung K (2017) 'Hypernudge': big data as a mode of regulation by design. Inf Commun Soc 20(1):118–136
Zarsky TZ (2019) Privacy and manipulation in the digital age. Theor Inquiries Law 20(1):157–188
Zuboff S (1988) In the age of the smart machine: the future of work and power. Basic Books
Zuboff S (2019) The age of surveillance capitalism: the fight for a human future at the new frontier of power. Profile Books