Intelligent Behaviour

How AI Transforms Behavioural Science, and Both Could Transform Our Lives

Stuart Mills

Unpublished Draft Manuscript

April 2025

Preface

This book started as something else. In the summer of 2024, I was looking for a project. The circumstances of the time meant I did not want to start anything new. I also knew, for the time being, I wanted to work on something alone.

In November 2022, I had published a book. It, too, was about artificial intelligence and behavioural science. It was not and is not a work I am especially proud of. There are various reasons for this.

The first is that 2022 had also been a difficult year, though for much different reasons. I had much on my plate, with a contract to write a book looming over me. I returned to one of my worse habits—sitting down and spewing out words. The second is a difficult relationship with my editor. By July 2022, I had submitted an initial draft. This draft was much longer than the wordcount I had been given (which I regarded as a positive from a consumer perspective), and the language insufficiently 'commercial.' My editor immediately requested edits, which amounted to the loss of thousands of words, including tangential details which, nevertheless, I felt added to the whole. The third was my loss of faith in the project. The revised manuscript was then sent for anonymous review, and these reviews were highly critical of the book in light of these reviews. I suggested to my editor we cut our losses—I had neither the time nor the enthusiasm to substantially change the book. My editor, instead, decided to repackage the book into something quite different to what I had initially signed up for. And so, the book was released, and I was unhappy.

The fourth was the price point of the book upon release. I had initially agreed to write a book which would be sold at a retail price, rather than a higher academic price. The repackaging meant that this pricing point was abandoned. A price was set which was much higher than the quality of the work warranted, and—to be frank—I was embarrassed to promote it. To this day, I have only ever publicly promoted the book *once*. Whenever it is brought up, I tell people not to buy it. As above, I was not proud of this work.

The fifth was the time of release. In January 2023, I was giving a talk where I was praised for the timing of the book. In the same month of release—November 2022—OpenAI released ChatGPT, and the world went made for AI. *What incredible timing!* And I am sure it was—I still receive a royalty cheque which, while not huge, is not a *negligible* surprise, either. The drawback, though, was that the book said *nothing* about generative AI. Indeed, scanning the pages, the book feels disconnected from the immediate reality of its subject. Like someone buying shares in

Lehmann Brothers in 2008—engaged, but looking in the wrong direction. Again: *embarrassing*.

So, back to the summer of 2024. I was looking for a project, one that already had some foundation and one which I could work on alone. I am not quite sure *when* it started, but at some point, I found the old .docx file for my first book. My plan had three parts. Firstly, I would edit and update the book, mostly to include discussions of generative AI, but also to generally make it better. I wanted to do the work I felt unable to do in 2022. I wanted to not have this embarrassing, incomplete work hanging over my head. I wanted to show I could do better. Secondly, I would give the 'revised' version away for free. I would upload the .pdf to my website, and make an ebook version available to download, too. It would not be a *convenient* buying experience, but I could not bring myself to ask for money anymore. I wanted to make those who had bought the original book whole, by giving them a product which might actually have been worth buying (still, though, not for the academic price). Thirdly, I would explain what had happened. This preface is that explanation.

I recall quite quickly that a 'revision' was not going to happen. Maybe it is the academic tendency to write too much. Maybe it was the shittiness of the starting material. Regardless, by around page 2 of the 'revision' I found I was cutting far more material than I had planned and adding *even more* new material to replace it. I cut out a whole chapter and planned to rewrite several others. Those rewrites then became, I guess, *writes*. By around page 5, I abandoned the original draft entirely. On a 'fresh' .docx file, I began to write. And at the top of the page, I called the project *Intelligent Behaviour*.

*

Colleagues and friends know that I hold a great affinity for Herbert Simon. In sharing my passion for Simon's work, I have come to realise I am not wholly alone in my affections. He is my favourite economist, though I do not love him for his economics. Rather, I find him and his work fascinating for its eclecticism, or at least, *apparent* eclecticism. Starting as a political scientist focused on organisations, critiquing economics to the point of becoming an economist, and then shifting to study artificial intelligence; Simon was a man who followed what was interesting given what was available, and in turn charted an intellectual path which, to an outsider, seems a bit mad.

Of course, it was not *totally mad*. One of the fascinating things about Simon's work, read broadly, is how disparate ideas seem to connect. Whenever Simon wrote, *whatever Simon wrote*, it seemed to always be about the same thing, lingering under the surface. I could perhaps say he was writing about *intelligent behaviour*—a not unreasonable statement, but one which I do not think I have the arrogance to claim. It is probably better to say he was writing about *effective* behaviour or *adaptive* behaviour—*how do thinking things survive?* Whether concerns with organisations or people or artificial intelligence, this is the question which links them. It is the question, too, which motivates the only piece of fiction Simon ever wrote, which he felt worthwhile to publish in his autobiography.

Simon played with a word for a time—*isomorph*—which I think is important to understanding his work. Two things are isomorphic when they are the same. Consider two congruent triangles, one twice as large as the other. These triangles are isomorphic, despite the difference in size, because their dimensions are the same, just subject to a different (arbitrary) scaling factor. Isomorph is one term; an (ironically) similar term is *homomorph*. Two things are homomorphic when they are similar. For instance, two triangles are homomorphic because they are triangles.

Simon used these terms—isomorphic and homomorphic—to describe games and problems in his work on human-problem solving. If two problems were the same, or similar, and one knew the solution to one (simple) problem, one thus had a means of solving the other (difficult) problem. Such would be a tremendously useful tool for an organisation seeking to conserve resources, a person seeking to make decisions quickly, or an AI system constrained by computing power. Thus, we have a connection, and in doing so, one might even claim (as I do in *this* book) that people, organisations, and AI systems, could all be understood as loosely *homomorphic*.

I bring up Simon because I am a fan, obviously. But I also bring him up to address (what at least to me is) an elephant in the room. The title *Intelligent Behaviour* was swirling around in my brain for a while as I contemplated writing a wholly new book. When I took that leap, I used that phrase, first planning and later simply allowing myself to change it at some point. As the writing progressed, I found myself liking the name more and more. When I re-re-edited the introduction (to finally remove the last traces of the earlier book), I did so with this title in mind, addressing the subject of *intelligent behaviour* directly. Yet, it was perhaps only halfway through the book I realised that my *Intelligent Behaviour* was a subconscious titular rip-off of Simon's *Administrative Behavior*. Of course, these books are different, and his is much better.

*

As summer turned to winter, I felt I could complete the draft before Christmas 2024. I did not. As a prime opportunity to stress-test this book arose in January 2025, I gave myself another month and produced the draft in time (this is the last thing I have to write). That is when the book was finally 'finished,' though I have learnt from past experiences that one should always make time for *one more edit*.

Unlike the experience of 2022, writing *Intelligent Behaviour* has *actually* been quite enjoyable. I have always known that one should write a book in papers, just as one does not *start* by running a marathon. But I have experienced the benefit of this in this work. While no chapter is just a rehash of one of my papers, each chapter organises my work of previous years into something that feels *homomorphic*, even if I did not plan things that way at the time. It has given me more of an appreciation for scholarly work, for long-form writing, for the process in general. It has allowed me to recognise, in, say, Simon's bibliography, how the pieces might have emerged, before there was any picture to put together. And it reminds me of a piece of advice, given to me by a valued colleague several years ago: a good thinker is a person whose work you can look at and go *yeah, I know what they're about*.

I am not confident I am a 'good thinker' in the way I believe Simon was a 'good thinker.' But I do know, unlike in 2022, I am not embarrassed of this work. There is always time for one more edit, but if there were not, I believe what follows is *not too bad*.

Chapter [0]—Introduction

> "Solving a problem simply means representing it so as to make the solution transparent."
>
> —Herbert A. Simon, *The Sciences of the Artificial* (1981, p. 153)

Questions and Answers

There are many discussions which could fall under the umbrella of, 'intelligent behaviour' and the topics contained therein, namely, behavioural science and artificial intelligence (AI). This book cannot possibly cover them all. Both the term 'artificial intelligence' and the term 'behavioural science' are too broad. They have been made broader still by boosters of various stripes, some acting in better faith than others. To adequately comment upon all features which might arise from the combination of these terms and the abstract notion of intelligent behaviour is impossible.

To offer commentary on a smaller but—hopefully—important set of ideas is more feasible. Therefore, I have restricted the focus of this book to a relatively narrow view of intelligent behaviour: *intelligent behaviour is the process of selecting options predicted to achieve goals*. Some readers might now choose to put this book down, having been satisfied with a workable definition of the titular phenomenon. It is my hope that most will not. From this simple definition spirals a whole world of questions which colour the subject matter and transform 'intelligent behaviour' from a cute phrase into something one can actually *use*, perhaps *abuse*, and certainly *critique*.

As above, intelligent behaviour is approached throughout this book via a behavioural science lens and is motivated by the question of how AI transforms the *what* and *how* of intelligent behaviour, when viewed through this lens. In brief, AI systems can be used to influence individual behaviours by analysing and altering the decisional contexts (or *choice architecture*) in which choices are made, and by extension, behaviours exhibited. There are many ways in which AI technologies may be applied to decision-making. There are also many angles to consider regarding *if* and *when* AI should be used in this manner. Such nuances all fall under the category of 'intelligent behaviour,' and are thus considered in this book.

I have also tried to write a book that helps a reader think through their own ideas within this rapidly developing area, which one might broadly call *behavioural technology*. It is too easy at present to be swept up in a wave of optimism about AI, as many have been swept up in a wave of optimism about the transformative powers of modern behavioural science in the past two decades or so.[1] Yet, *candidly*, it is not really that much more difficult to be sceptical about both fields, either. It is quite easy to hand wave away this or that result, this or that technology; to proclaim that the future is not coming, while desperately not looking at the hands turning upon the face of one's watch. Whatever is to become of AI and behavioural science, separately and in combination, these disciplines exist, and thus there is some consequence to which thought must be given.

What this book aims to achieve is *realism* when evaluating what intelligent behaviour might mean for people and society today. Herbert Simon—whose ghost haunts the pages of this book—once argued that our obligation to one another is to answer some questions, but in doing so, to leave more interesting questions for others to someday answer. I am not sure that this book poses many answers, but I do hope that the ideas herein contained equip one with some means of asking, and perhaps answering, the various interesting questions which emerge.

A Meeting of Minds

To provide such equipment, one of my tasks in this book must be to translate various ideas from one discipline into the language of another. But what is interesting in this regard is how much behavioural science and artificial intelligence are already related. In many ways, the fields of AI and behavioural science are two sides of the same coin that is *intelligent behaviour*. While AI researchers endeavour to, "create mind in machines" (Turkle, 1988, p. 244), the objective of behavioural science (and the fields which constitute it, namely psychology and—to a lesser extent—economics) is to explain the product of the mind (i.e., behaviour) within the world (Haig, 2014). The similarities between fields, therefore, originate from the central objective of probing the 'black box' of intellect and reason (and thus *behaviour*) which is possessed by the human, and is given to the machine.

These disciplines are also related philosophically, following what Skidelsky (2023, p. 9) has called the "mechanical philosophy."[2] With the emergence of industrial machinery in the late eighteenth and early nineteenth century, interesting questions

[1] And to which many are now discovering that such transformations are often much more modest. See, for instance, Chater and Loewenstein (2023).

[2] This may also be associated with the philosophy of *mechanism*—a philosophical school which contends that all things can be reduced to fundamental components which can definitely be understood. In essence, that all phenomena can be engineered.

began to be asked about the nature of human beings, and the two (simplified) conclusions which came from these inquiries constitute the mechanical philosophy which has bearing on AI development (also see Joque, 2022; Pasquinelli, 2023) and on what I will call 'modern behavioural science.'[3] Firstly, that "the scientific method… enable[s] laws of human behaviour to be established, just like the laws of physics" (Skidelsky, 2023, p. 9) and secondly, that through these 'laws' the mind, body, and perhaps *soul*, could be created within machines (Scriven, 1953; Turkle, 2004, 1988).

It is thus helpful, as an introductory discussion, to briefly consider the entangled history of AI and behavioural science (Miller, 2003; Turkle, 1988), which I will tell as a 'co-history' of the disciplines. By a co-history, I mean two histories which can interact and influence one another, but which can also be viewed as moving independently. Interactions may be accidental; overlaps overlooked; and agendas driven by different economic, intellectual, and political concerns.

A simplified co-history could begin in several places—I will start with economics.[4] Economics is a field concerned with optimisation under human wants and needs (i.e., preferences), and is thus a specialised field studying human behaviour (loath as some economists might be to admit this). Von Neumann and Morgenstern's (1944) *Theory of Games and Economic Behaviour* is possibly the most famous attempt to *axiomatically* describe rational human behaviour (i.e., the 'rules' of rationality, or, perhaps, the 'laws' of human behaviour), and thus propose a model of intelligent behaviour.[5] Such axioms proposed a model of human behaviour

[3] By *modern* behavioural science, I mean specifically the popular applied branch of behavioural science that has emerged in the twenty-first century, following Kahneman and Tversky's 'heuristics and biases' programme in the late twentieth century. The notion of 'systematic biases' which emerges from this programme underpins many applied efforts (e.g., Thaler and Sunstein, 2008) and is philosophically quite similar to the notion of "laws of human behaviour." Indeed, it is through appealing to these 'laws' that Herbert Simon's notion of *bounded rationality* likely found a (somewhat stultified) place within the modern behavioural science programme (see Petracca, 2021).

[4] Another viable starting point would be Norbert Wiener's work on cybernetics in the 1940s. Indeed, much modern social science can be traced back to the work being undertaken by the U.S. military during the Second World War, and by the RAND corporation, thereafter, including work on AI, behavioural science, and the study of intelligent (system) behaviour more generally.

[5] Prior to von Neumann and Morgenstern's (1944) work, the less mathematical and somewhat more erratic discipline of political economy had frequent musings on human behaviour as diverging from mathematical formalism. For instance, see Keynes' (2017) various theories, including *liquidity preference, animal spirits,* and *behaviour under uncertainty*, all of which can retrospectively be understood as a kind of behavioural economics (Thaler, 2015). Equally, as Skidelsky (2023) notes with reference to Carlyle's 1829 work, and Pasquinelli (2023) notes with reference to Babbage's 1832 work, the changing technological landscape of the Industrial Revolution was catalysing new ideas about where people fit into a mechanical world, and in turn, encouraged perspectives of "humans… [as] 'wired up' parts

where a person would, for instance, be consistent in their preferences, or would maximise their expected utility.

For AI research, this model of behaviour was a worthwhile launching pad for thinking about how to make a machine intelligent (Russell, 1997). Work on computer chess in the Soviet Union, for instance, cross-fertilised with efforts to network the economy and automate economic activities according to rational economic planning (Peters, 2017). Herbert Simon (1981, 1956, 1955) would provide powerful interventions which critiqued and further developed the axiomatic approach to AI development, supported in part by Miller's (1956) work on human cognitive limitations (Newell and Simon, 2019).

Simon's specific idea—*bounded rationality*—had a profound impact on what AI was expected to be (Russell, 1997). Bounded rationality holds that decisions (to choose A, or to choose B; to go left, or to go right) must be made under conditions of limited information processing power (Simon, 1955) and under the influence of environmental factors (Simon, 1956). A person cannot evaluate all relevant information when choosing to buy a red car, or a blue one; an organisation cannot evaluate all relevant information when deciding whether to build a new factory or refurbish an old one; and a computer cannot evaluate all relevant information when calculating whether moving the knight is better than moving the bishop (Simon, 1981). For Simon, all problem-solvers must employ *heuristics* to cut through the processing impediments we all face and reach a decision which allows us to move on with our lives (Newell and Simon, 2019)—a much more realistic, if less ambitious, view of intelligent behaviour.[6]

For much of the twentieth century, computers had extremely limited information processing power, even if they exceeded the power of humans to crunch through some problems. Many AI

of a complex technological system" (Skidelsky, 2023, p. 9) leading to a drift towards greater formalism and 'rational' human behaviour. Simon (1997b, p. 16), too, notes this tendency within economics, commenting that since Adam Smith's *The Wealth of Nations*, economics has engaged in a process of increasing the necessary rational performance of humans, and moving away from the "rationality of everyday common sense."

[6] There are two relevant notes here. *Firstly*, Simon's notion of heuristics says little about whether choices are 'good' or 'bad.' For Simon, the universal bad is indecision, for this forestalls all effective exercise and means a decision-maker will, inevitably, fail to adapt to their environment. The notion of *satisficing*—of choosing a 'good enough' option—is generally grounded in the notion that choosing some option will be better than being stuck trying to choose the best one. *Secondly*, an astute reader will notice the emphasis on information processing, but not on environmental factors. As Petracca (2021) has argued, this omission was convenient for many parties seeking to utilise the idea of bounded rationality in AI, in economics, and in behavioural science, and omission of environmental factors was even *encouraged* by Simon insofar as it gave bounded rationality longevity as an idea (Simon, 1996).

problems—like computer chess—demanded more processing power than was available given the technology of the time. As such, bounded rationality offered an interesting conceptual way of thinking about problem solving and developing effective tools for reducing the computational complexity of various problems.[7] This was an important idea which shaped the development of early 'symbolic' AI.[8]

Symbolic AI systems follow logical rules which allow one to determine what action one should take, given some input. If effective rule-based shortcuts—or *heuristics*—could be identified, computationally-taxing processes may be greatly reduced, and thus the effective usefulness of computers (in organisation, in administration, and in everyday life) would substantially increase. Simon's efforts bore some fruits: at the famous 1956 Dartmouth Conference on AI, Simon and his collaborator Allen Newell were the *only* attendees to have an actual computer programme to demonstrate (Simon, 1996).[9]

Bounded rationality can also be seen as foundational to the critique of the axiomatic approach as a *description* of human behaviour given by Tversky and Kahneman (1979, 1974, 1973)

[7] For instance, Simon was the academic 'grandfather' of Hans Berliner (via Allen Newell), the famed MIT professor who championed the computer chess problem and inspired the team that would eventually develop Deep Blue, the computer that beat the chess grandmaster Garry Kasparov.

[8] Symbolic AI is designed to follow various logical rules depending on the input it is given. For instance: IF $x = 1$, choose A; ELSE, choose B. All rules can be described in terms of a *systems of symbols* (e.g., mathematical symbols, logical operators, musical notation, chess notation), hence the notion of AI being symbolic.

Symbolic AI has some advantages. For instance, it is relatively easy to understand what the AI is doing at any given moment. It also has several disadvantages, primarily that not every task is easily described symbolically. Minsky and Papert (2017) illustrate this with what one could call the cat problem. Using any set of symbols, and as many rules or logical statements as one wishes, derive a set of statements which accurately identify a cat, while not erroneously labelling non-cat entities. With enough time, and enough rules (and perhaps enough knowledge of cats), one may be able to accomplish the task. But as the set of statements must be so specific, such a task is likely impossible for any intelligent entity to complete. This problem is also nicely summarised in this Wittgensteinian joke: *A man walks into a bar. The barman asks him, "what is a horse?" The man replies, "a horse has four legs, and you can sit on it." The barman looks shocked. "Who let all these horses in here!?"*

As discussed below, perceptron-based AI (neural AI) resolves this epistemic problem ('what is a cat?') by never actually articulating what a cat is, but merely identifying opaque statistical relationships between entities labelled as cats and non-cats.

[9] This programme used heuristic-based search to prove logical statements taken from Bertrand Rusell's famous work attempting to establish a formal, logical basis for all of mathematics. Russell, apparently, was impressed. When developed, Simon jubilantly announced to his class in January 1956 that he and Allen Newell had created a "thinking machine" over the Christmas break—quite the vacation project!

within cognitive psychology.[10] For several decades in the twentieth century, psychology had been dominated by *behaviourism*, a school of thought which contends that internal mental states are unnecessary to understand behaviour. Instead, one must rely on what can be observed, focusing on what stimuli will produce desired behaviours (Skinner, 1984).

The cognitive psychology approach took no less of a view than the behaviourists that the mind was a 'black box' which needed to be investigated through systematic variance of stimuli and scientific observation of responses (Haig, 2014; Skinner, 1976; Turkle, 1988). But unlike behaviourism, the emerging cognitive psychology to which Tversky and Kahneman were a part contended that understanding the mental mechanisms taking place within the 'black box' was crucial to understanding human behaviour.[11] One could not just rely on observed stimuli, but instead needed to construct theories of cognitive processes. It was in trying to formulate explanations for how people choose between options that Tversky and Kahneman found the axioms of

[10] I have chosen my language carefully, here. As Heukelom (2012) notes, Tversky and Kahneman may have been unaware of Simon or bounded rationality at the time they were developing their 'heuristics and biases' programme, and indeed, the resurrection (in a manner of speaking) of bounded rationality within modern behavioural science has been one in support of, and wrapped in the language of, the heuristics and biases programme, rather than the other way around (Gigerenzer, 2007). As Heukelom (2012) has argued, this does leave some important differences between these perspectives to be missed. Notably, that Simon's interest concerns *how people decide* (i.e., how people navigate everyday decisions, hence Simon's use of the phrase 'problem solving' rather than 'decision-making'; see Simon, 1996), whereas Tversky and Kahneman's interest concerns *how people simplify to be able to decide* (i.e., how people make singular decisions for which preferences and diverges from said preferences can be measured, hence modern behavioural science's focus on singular decision-making rather than everyday behaviour). Petracca (2021), too, has contributed some interesting scholarship on this question, and offers some worthwhile answers related to Simon's motivations around legacy.

[11] Following Turkle's (1988) account, part of what 'killed' behaviourism was the invention of the computer. The argument goes that the computer, on the face of it, was a behaviourist machine, insofar as one could—just as with humans and animals—give the computer inputs (stimuli) and observe outputs (responses/behaviours). But unlike humans and animals, computer scientists could tinker with the internal structure of the computer, demonstrating the importance of internal states in determining *how* stimuli are received, and *how* responses are exhibited. This, in turn, struck a deathblow to behaviourism's minimisation of internal mental states.

This is likely too simple of an explanation. *Firstly*, scholars such as Chomsky (1959) are often held to have contributed to the decline of behaviourism. Chomsky's (1959) theory of universal grammar suggested there may be innate properties to human behaviour and intelligence, which contradicted behaviourist ideas about reinforcement learning. *Secondly*, behaviourist approaches were better suited for relatively simple, controlled environments, and as applied psychology became involved in more complex arenas (e.g., factory management), behaviourist theories failed to provide satisfactory results, prompting a shift in intellectual energies given the declining intellectual booty to be won.

economic rationality failed to describe observed human behaviour (Lewis, 2016).

Instead, what Tversky and Kahneman began to uncover were a series of 'systematic biases' in human decision-making. They found that people tended to over-estimate the likelihood of events which they could readily bring to mind (availability bias; Tversky and Kahneman, 1973), that preferences for risk would switch depending on whether choices incurred losses or gains (loss aversion; Kahneman and Tversky, 1979), and that people would choose options which contradicted objective probability when those options appealed to real-world representations (the conjunction fallacy; Tversky and Kahneman, 1983).

Such observations did not align with von Neumann and Morgenstern's (1944) axiomatic view of human behaviour (or those axioms which were to later develop), but they did align more closely with a bounded rationality perspective on decision-making. For instance, it is difficult to evaluate all information which may be relevant when estimating the likelihood of a natural disaster. But it is relatively easy to recall an instance a couple of years ago when your basement flooded (the availability heuristic). Such information may therefore offer an easy means of reaching a decision (e.g., that flooding is more likely than a wildfire).[12]

By the 1990s, studies into human biases and decision-making were producing an impressive body of work which professed to provide a deeper understanding of human behaviour than economic descriptions had up to that point (Dhami and Sunstein, 2023). *Anomalies*, written by Kahneman, Knetsch and Thaler (1991), would stand as an important attempt to combine key ideas from this intellectual revolution, including loss aversion (Kahneman and Tversky, 1979), the endowment effect (Knetsch, 1989), and the status quo bias (Samuelson and Zeckhauser, 1988).[13]

[12] As above, there is an important difference between the way Kahneman and Tversky utilised the term 'heuristic' and the way it was used by Simon, owing to their intellectual disconnect and the independent adoption of the term (Heukelom, 2012). Tversky and Kahneman (1974) understood the heuristic as being a way of *simplifying* a complex problem. One may not know the likelihood of Gotham flooding, but one may be subjectively surer of the likelihood of one's basement flooding owing to past experiences. Ergo, the question is simplified from 'the likelihood of Gotham flooding' to 'the likelihood of flooding *somewhere*.' For Simon, heuristics are to be used as a means of deciding in a satisficing way. For instance, if Gotham is next to a river, rather than in a desert, one could conclude, '*there is a reasonable chance that flooding could happen in Gotham*,' and thus predict flooding to be more likely than a wildfire. The latter, if it simplifies, simplifies the *evaluation* criteria (from 'what is the likelihood of flooding' to 'is flooding more likely than a wildfire?'), whereas the former, as above, simplifies the problem itself.

[13] Technically, *Anomalies* was an ongoing series by Thaler documenting some behavioural economic results. However, the entry by Kahneman, Knetsch and Thaler is the most famous.

Further development of these ideas would occur throughout the 1990s,[14] resulting in various investigations into the *applications* of these theories of human behaviour towards *behaviour change* (e.g., Johnson and Goldstein, 2003; Madrian and Shea, 2001; Thaler and Sunstein, 2003). These efforts culminated in Thaler and Sunstein's (2008) book *Nudge*, which suggested the context in which humans make decisions (so-called *choice architecture*) could be actively redesigned, drawing on prior behavioural theories, to positively influence the behaviour of decision-makers. The suggestion was that such redesigns could support people in choosing options predicted to achieving their goals—in a manner of speaking, to support intelligent behaviour. Modern behavioural science had arrived.

The path from bounded rationality was less smooth for AI development. Progress in symbolic AI slowed in the 1960s and 1970s, culminating in the so-called 'AI winter' (Russell, 2019).[15] The challenges of specifying effective rules for AI systems to reason through problems, coupled with continual computational processing difficulties, would eventually push AI development away from the symbolic approach, and towards an alternative 'neural' approach which continues to dominate today (Russell and Norvig, 2020; Turkle, 1988).

Neural AI, including *neural networks* and *deep learning*, are systems loosely inspired by the structure of the human brain and built from machines called *perceptrons* (Minsky and Papert, 2017). Perceptrons take in large amounts of data; 'combine' these data using weights; and make predictions based on this weighted calculation. These predictions can then be compared with known answers, before the weights are adjusted to improve the accuracy of future predictions. This approach builds off ideas of connectionism in psychology (e.g., Hayek, 1999; Rosenblatt, 1962) and earlier theories of neurone activation in the human brain (von Neumann, 2000), whereby connections become stronger through practice and repeated exposure to similar stimuli.[16] The perspective also draws on some theories of feedback found in the cybernetics literature (Medina, 2014; Wiener, 1950).

[14] For instance, Tversky and Kahneman (1992) would revise and update their original prospect theory, which derived loss aversion, while Benartzi and Thaler (1995) would incorporate a temporal dimension into the theory.

[15] The notion of an 'AI winter' has been disputed by writers such as Haigh (2023), who argue that the 1970s represented more of a transition in approaches to AI than a full-scale abandonment or loss of faith in the discipline.

[16] Neural networks consist of several (many) perceptrons, while deep learning describes a neural network with several (many) hidden 'layers'—essentially, a big neural network. These AI systems are currently the most successful AI approaches for tasks such as identifying cats from dogs or beating humans at games such as *Jeopardy* and *Go*. They are also the essential building blocks to all large language models (LLMs) and generative AI systems currently developed.

Quite what the weights contained within neural AI systems actually represent is difficult to describe, and this is an ongoing conceptual (and increasingly regulatory and political) challenge for AI practitioners (Garcez and Lamb, 2020; Garnelo and Shanahan, 2019). Though, as Pasquinelli (2023) has argued, the neural approach has important advantages in terms of requisite knowledge. When one trains a neural AI system to distinguish a dog from a cat, one needs to know *nothing* about either entity. One thus substitutes knowledge with its more abstract representations (e.g., data) in the hopes that a model can be constructed which *simulates* knowledgeable behaviours.[17]

Regardless, both symbolic and neural AI approaches draw on the 'mechanical philosophy' which suggests that behaviour and the knowledge involved in such behaviours can be disassembled (*divided*) and systematically arranged (*rationalised*) to recreate behaviour in machines. AI algorithms, whether disentangled in the form of symbolic AI or entangled in the form of neural AI, represent attempts to codify the 'laws of human behaviour' and translate them to machines (Pasquinelli, 2023). But so too, one might argue, does the emergence of modern behavioural science, with its efforts to describe behaviour in terms of 'systematic biases,' and 'simplifying heuristics.' These ideas, in turn, point to 'choice architecture,' which can be proactively designed and redesigned (Mills and Sætra, 2024a).

The benefit of exploring this co-history, besides equipping a reader with some relevant background ideas, is to recognise that the collision of AI and behavioural science is not the meeting of strangers, but the *reunion* of estranged cousins as both seek to understand intelligent behaviour. It emphasises that while one might consider the *new* possibilities that AI presents for behavioural science—and this book certainly will do this—one should also consider that one use of AI in behavioural science can (and *should*) be as a mirror to integrate the intellectual and philosophical developments which have evolved behavioural science into its modern incarnation. Perhaps in doing so, one may catch a glimpse

[17] A helpful, though incomplete, way of thinking about this process is to imagine that a neural AI predicts a real phenomenon (R) from data captured about the phenomenon (D), such that $R = \omega D$. Data cannot perfectly describe something that is real, just as a map can never be a one-to-one scale of wherever it maps (Simon, 1981). Something is inevitably lost in the process of rendering reality into data, and thus something must be added in the process of using data to predict reality. Neural AI models, through prediction and feedback, might be said to home in on a precise value of ω such that its predictions are reasonably accurate. This value has no inherent meaning. At best, it captures only a statistical relationship between data and observation. It is a brute-force, data-driven means of filling in a gap in our ability to describe reality—essentially, it is a way of solving the cat problem. Hence, there is no 'knowledge' of reality, only a simulation of knowing.

as to where the discipline could go next. Considering both, one arrives at an important set of ideas concerning intelligent behaviour.

The Structure of This Book

I am now at the point where it seems appropriate to really begin this book. Chapter 1 introduces various definitions which will establish the language framework upon which the rest of this book is built. It involves distinguishing between predictive AI and generative AI systems; an important division given the behavioural implications of these functionally different AI applications.

Chapter 2 focuses on personalisation within behavioural science. Personalisation has long been an ambition of modern behavioural science. With the advent of AI as a means of analysing data and predicting optimal designs, the promise of personalisation may now be realised. The chapter outlines what choice architecture is; why personalisation offers theoretical benefits for behavioural science; how approaches to personalised design can be understood as an algorithm; and the role of personalised paternalism.

Chapter 3 focuses on emerging applications of AI within behavioural science beyond personalisation. It considers how AI systems may be incorporated into behavioural interventions to directly affect people's behaviour, such as via chatbots. It also considers how these technologies can further behavioural science knowledge through the development of new methodologies, such as simulation modelling.

Chapter 4 moves beyond questions of how AI can be used in the study and influence of behaviour, and towards questions of how *ought* AI be used, and how is it *actually* used. Through an exploration of recent arguments for using AI systems as decision-making aids, in public, personal, and professional life, this chapter emphasises the wider politics of both AI and behavioural science.

Chapter 5 continues the exploration of *ought* and *actually*. However, rather than exploring the use of AI as a tool, it considers the implications of AI being a *machine*; an autonomous system with the authority to influence people. This chapter introduces the idea of the *autonomous choice architect* and emphasises how technological development strains the ethical arguments, developed by behavioural scientists, for the use of behavioural science.

Chapter 6 concludes.

Chapter [1]—Definitions

> "[I]t is, with our habits of thought and of expressing thought, very difficult to express any truly complicated situation without having recourse to formulae and numbers."
>
> —John von Neumann, *The Computer and the Brain* (2000, p. 74)

Why Definitions Matter

Why devote a whole chapter to definitions? This book has already begun with a definition. Why add more? If one wanted to read a bunch of definitions, one could presumably start thumbing through the dictionary.

As a pedant, I have always felt definitions matter. When I began researching AI, and talking to those researching similar areas, I quickly felt the rush of pedantry come over me. AI is notorious for its loose definition (Russell, 1997).[18] It is another attribute shared with behavioural science. One review of the literature in 2020 found no less than *fifty-five* definitions of artificial intelligence (Samoili *et al.*, 2020). No doubt that number has grown. It should go without saying, but an excess of definition implies a lack of coherence.[19]

When trying to "create mind in machine" (Marvin Minsky's definition of AI, according to Turkle, 1988, p. 244), one must be careful one understands precisely what this means (as one Google engineer who felt his model had become sentient discovered; de Cosmo, 2022). *Socially*, it also matters that one knows what one is talking about. AI is being used to direct sophisticated weapons systems (Russell, 2023); to decide who should receive assistance from the state, and who is to be left to fend for themselves (Katwala, 2020; Kinchin and Mougouei, 2022); and its supply chains are threatening the environmental integrity of several communities (Hogan, 2015; Lehuedé, 2024). As one will see in this book, AI is increasingly being used to shape everyday human

[18] Russell (1997, p. 57): "AI is a field whose ultimate goal has often been somewhat ill-defined and subject to dispute. Some researchers aim to emulate human cognition, others aim at the creation of intelligence without concern for human characteristics, and still others aim to create useful artifacts without concern for abstract notions of intelligence."

[19] The term 'artificial intelligence,' so the story goes, only came about through a clash of personalities, as Warren McCulloch—a father of AI—disliked Norbert Wiener and wanted to distance his work from Wiener's work on *cybernetics*, a field which *also suffers from over definition* (Medina, 2014).

behaviour, too. Poor definitions make for poor discussions, applications, and analyses, all of which must be avoided when the stakes are so high.[20]

My objective with this chapter is to offer some definitions of concepts that will be fundamental to the remainder of this book. Given definition can become a bit tiresome, I will endeavour for brevity, and leaving unnecessary comments to the footnotes. I will begin with some more fundamental definitions—of *behaviour*, of *intelligence* (leading to a restatement of *intelligent behaviour*, as above), and of *machine*—before moving onto defining two distinct types of AI, predictive AI and generative AI.

Behaviour, Intelligence, and Machine

Behaviours are the actions we take in the world.

For most uses of the word, this definition will suffice. Though, it could also be accused of being too broad, which, as above, can be unhelpful when it comes to definition. For this reason, Furr (2009, p. 372) defines behaviour as "verbal utterances or movements that are potentially available to careful observers using normal sensory processes." This definition is offered as it, "excludes many internal physiological responses such as neural events and blood pressure… [as well as] many external physiological responses such as blushing or sweating."

This is quite reasonable, though the advent of new technologies means that increasingly those 'actions' which come to constitute one's understanding of behaviour have grown (Rauthmann, 2020). fMRI machines, for instance, can observe neural events, while mobile phones can now track a person's eyes and subtle facial expressions (Valliappan *et al.*, 2020). AI—as an analytical technology—makes these data more useful, while the diffusion of technologies such as smart phones throughout society make these data more readily accessible, and the insights of these data more immediately actionable. To restrict behaviour to 'normal sensory processes' when the *extra*ordinary is increasingly rendered mundane by digital technologies is likely a mistake.[21] As such, an

[20] A reader should note that I am not aligning myself with several 'AI safety' and 'X-risk' perspectives which imagine some kind of superintelligence threatening to destroy the world (e.g., Bostrom, 2014; Moynihan, 2020; Tegmark, 2017). In my opinion, such arguments are fantastical, motivated in part by too much science fiction and a good dose of political cynicism. The best description of much of the AI apocalypse discussion is that it is *scareware*—a kind of malware which plays off human ignorance and fear to affect a behaviour one otherwise would not exhibit. People—and *investors*—take you seriously when you claim your product could destroy the world. A different marketing of one's product—one might say a more *honest* or *accurate* marketing—might elicit a less serious response.

[21] This is something of a Kuhnian perspective. Kuhn (2012) notes that technology does not simply emerge from our scientific knowledge; it emerges in

immediate, broad consequence, of AI within behavioural science is that the technology will explode the catalogue of phenomena which a behavioural scientist can measure and evaluate when studying human behaviour, as discussed further in Chapter 3 (e.g., Buyalskaya *et al.*, 2023; Mills, Costa and Sunstein., 2023).

It is also interesting to see that Furr's (2009) definition talks of behaviours which are "potentially available" to be observed. This is worthwhile to highlight for two reasons. *Firstly*, when AI is used to predict human behaviour, the AI system does not necessarily choose a single behaviour, but rather calculates a *probability distribution* of behaviours which it has been programmed to consider viable or possible. In this sense, while aspects of Furr's definition benefit from reconsideration of new technologies, the notion that behaviours are actually those which are *potentially* available, rather than *necessarily manifest*, is very useful when thinking about AI. *Secondly*, because it gives license to thinking about *intelligent* behaviour specifically. For this adjective to mean anything, one must accept that there are alternative behaviours that, in a given moment, one could demonstrate. For one to make a mistake, one must have always had the opportunity to avoid that mistake; else, notions such as 'behavioural biases' and 'correcting biases' become meaningless. It is thus conceptually necessary to understand behaviour not as what is observed, but as what *could* be observed. As what is *potentially* available.

It is from here that one can arrive at a useful definition of intelligence. Russell (2019, p. 9) offers a succinct definition: "Humans are intelligent to the extent that our actions can be expected to achieve our objectives." From this, one readily arrives at the introductory definition of intelligent behaviour as behaviours which are taken so as to achieve one's goals. For instance, if one were to find oneself in a burning building, an intelligent behaviour would be to leave the building—assuming the goal is to survive the fire—while a less intelligent goal would be, say, to do a dance while the flames surround you. As above, recognising that behaviour can be thought of as a *set* of *potentially* available actions, one arrives at the full definition of intelligent behaviour given in the introduction—the *selection* of actions predicted to achieve one's goals.

In both psychology and AI studies, this goal-orientated understanding of intelligence, and intelligent behaviour, has been

response to what we *want* to know, and changes what we *can* know. Galileo needed to develop various optical technologies to then develop his ideas about the cosmos; the telescope thus changed what it meant to *know* the universe. Digital technologies like AI are having the same effect on those who study human behaviour. Whether this is widely appreciated at present is harder to discern.

adopted (e.g., Silver *et al.*, 2021; Sternberg, 1999). Sternberg (1999, p. 292), for instance, has championed the notion of "successful intelligence" in contrast to older ideas of a general factor of intelligence, such as IQ. When understood as selecting behaviours to achieve goals, measurements such as IQ become unhelpful, as they say essentially nothing about how a person actually survives, adapts, and *succeeds*, within the world (also see Simon, 1997a).[22] Sternberg's notion of successful intelligence is broader precisely because of these criticisms that intelligence cannot be divorced from context of the goals one seeks to fulfil. John McCarthy (2007, p. 2, emphasis added)—another 'father' of AI—has offered a definition which clearly acknowledges a similar multifaceted perspective: "Intelligence is the computational part of the ability to achieve goals in the world. *Varying kinds and degrees of intelligence* occur in people, many animals and some machines."

The multifaceted nature of intelligence can be taken further. For instance, Possati (2020) and Turkle (2004) have emphasised that what is considered intelligent changes depending on time and perspective. At one time, it was considered wise for a leader to forgive the debts of their subjects, so as to maintain social harmony; today, it is considered wise for a leader to apportion state resources to enforce creditor rights over the debts of citizens, less the leader's credibility (their 'bond') be called into question (Graeber, 2014). Relatedly, the so-called 'AI Effect' describes the phenomenon that whenever a computer is able to do something thought previously only the reserve of humans, that thing ceases to be a benchmark for intelligence (Haenlein and Kaplan, 2019). This, of course, denies the possibility of *artificial* intelligence, while the whole goal-orientated perspective suggests that intelligence is much more subjective than objective.[23]

From the perspective of modern behavioural science, this is important to appreciate, as it should prompt one to question not only if the behaviour change one is introducing is the helpful and ethical, but also whether *any* intervention to change behaviour is worthwhile.[24]

[22] While Simon (1997a) does not discuss IQ, the notion of adaptability and success within an environment, rather than in the abstract, is a clear theme of his thinking. In a criticism which should be relevant to modern behavioural science, Simon argues that what must be understood is not how one makes a single decision, but how one dynamically navigates from where they are to where they want to be (i.e., their goals) over many imperceptible decisions, and how feedback and new information leads to adaptions in one's choices. Fittingly, Simon calls such behaviour and decision-making *strategy*.

[23] Chapter 4 will explore this idea more completely, and with additional examples.

[24] This is not especially dissimilar from various critiques made by the 'fast and frugal' heuristics perspective, which has always been more aligned with the notion of adaptive behaviour and goal-oriented behaviour than the modern

Lastly, in a critique similar to that of Possati and Turkle, Pasquinelli (2023, p. 235) argues that "intelligence is a social process by constitution." This is to say, that intelligence is socially constructed based on interactions between people and the world. Pasquinelli's critique emerges directly from considering AI; he suggests that modern AI systems are not intelligent insofar as their "internal logic" (p. 234) contains some quantum of intelligence, but rather, these systems are intelligent insofar as they produce outputs which align with what *people* determine to be viable, useful outputs within the world that they, not the AI, inhabit. From the perspective of intelligence being goal-orientated (or value-based; Simon, 1997a), this should make intuitive sense: where else might one's broad goals come from, if not from one's social experiences?[25] Immediately, though, it is adequate to simply say that some understanding of intelligence as being goal-orientated and a function of interaction with the world is useful from a behavioural science perspective.

Finally, it is worth defining what is meant by the word *machine*. Compared to behaviour and intelligence, this might seem like an outlier. Nevertheless, its definition has implications later when considering the role AI could play in *applied* behavioural science, and the *absence* of human oversight therein enabled, as will be the focus of Chapter 5.

While the words 'tool' and 'machine' are often used interchangeably (Skidelsky, 2023), Marx (2013) emphasises an important difference between them;[26] a difference which Turkle (1988) and Gunkel (2017) have extended to discussions of AI and society. Marx (2013, p. 257) argues that a tool must be given its

behavioural science perspective. Per a previous footnote, again, Chapter 4 explores these ideas more fully.

[25] Note the distinction between 'broad' goals and goals generally. In the burning building example, one seeks to satisfy a more primal, biological goal of survival. This may only tenuously be chalked up to socialisation; it is a much more automatic response. Nevertheless, one who seeks to escape a burning building might still have broader, more abstract goals which they seek to fulfil. One might be pedantic and suggest that one pursues the immediate goal of survival *in order to* make it more likely one can attain a broader goal. As Simon (1997a) notes, goal-orientated behaviour can be quite arbitrary concerning precisely what a person's goal is at any given moment.

[26] Marx (2013, p. 256-7): "Mathematicians and mechanicians, and in this they are followed by a few English economists, call a tool a simple machine, and a machine a complex tool. They see no essential difference between them, and can even give the name of the machine to the simple mechanical powers, the lever, the inclined plane, the screw, the wedge, etc. As a matter of fact, every machine is a combination of those simple powers, no matter how they may be disguised. From the economic standpoint this explanation is worth nothing, because the historical element is wanting. Another explanation of the difference between tool and machine is that in the case of a tool, man is the motive power, while the motive power of a machine is something different from man, as, for instance, an animal, water, wind, and so on."

"motive power" by a person, whereas a machine possesses its own, independent, source of motive power. As such, a tool requires human action, whereas a machine—in principle—does not. This distinction creates the possibility of machines automating jobs and displacing people, which is a common feature of many discussions of AI today (e.g., Acemoglu and Johnson, 2023).

This distinction is relevant to the present discussion as it invites one to ask whether AI should be used by behavioural scientists as a tool, or whether the efforts of behavioural scientists can (and should) be used in the construction of a behavioural AI *machine* or *autonomous choice architect*: a machine that is built to influence human behaviour without necessarily requiring human input or oversight (Mills and Sætra, 2024). Without defining a machine as an entity possessive of its own motive power, such questions do not emerge. But such questions strike at the heart of much of modern behavioural science, especially given its emphasis on maintaining freedom of choice while supporting people to make 'better' choices (e.g., Thaler and Sunstein, 2003). As above, such questions will be dealt with in Chapter 5.[27]

Predictive AI and Generative AI

Much of the interest in AI in recent years has been driven by the emergence of generative AI technologies. Behaviourally, this is hardly surprising. *Salience* captures how noticeable or attention-grabbing something is, and generative AI technologies—by virtue of their outputs—tend to be very salient. They have thus dominated recent conversations around AI systems.

One interesting aspect of this dominance is that it is not reflected in prevalence. The use of AI has grown steadily over the past ten to fifteen years, while generative AI products have only really seen mainstream adoption since around 2022. As such, most of the AI technology that each of us encounters in our daily lives is *not* generative AI, but predictive AI.[28] This statement extends

[27] These questions are furnished further when one considers, once again, the notion of mechanical philosophy. For Skidelsky (2023, p. 9), our present era is not exceptional because machines exist, but because in past epochs, "machines did not determine the conditions of life." Skidelsky suggests that today, "we live in a machine age," where "humans are 'wired up' parts of a complex technological system," which leads to, as noted in the previous chapter, attempts to apply mechanical logics to human experiences, and to expound upon "laws of human behaviour." Interested readers should also consider the commentaries of the social critic Ivan Illich, especially Illich's (1973) *Tools for Conviviality*. Chapter 5 elaborates on this perspective with some discussion of automation bias and the notion of *quasi*-motive power.

[28] From a technical perspective, all generative AI is just predictive AI, but, in a manner of speaking, 'run backwards.' This diffusion approach is fascinating and warrants its own public praise. Still, the distinction drawn here largely focuses on the social, rather than technical, distinct between generative and predictive AI.

even to those people who, for instance, use an application such as ChatGPT every day. Even for this enthusiastic adopter, predictive AI systems will dominate their experiences with AI. This oversight is not due to some pernicious reason. Quite simply, predictive AI technologies are less salient than the generative variety.

This observation is reflected in the technological focus of this book. It makes sense to apportion most attention to the type of AI technology—predictive AI—that is most prevalent and will likely continue to be most prevalent, even if not in popular discussions.[29] As such, Chapters 2, 4, and 5 are mostly concerned with predictive AI. Chapter 3 gives more attention to generative AI, though no chapter is subject to a strict demarcation between the two. What distinguishes predictive AI from generative AI?

Predictive AI utilises machine learning technologies to build models capable of predicting outcomes from a set of possible options in a manner indicative of intelligent behaviour. These predictions may then be used by people to inform decisions, such as by policymakers when tweaking policy parameters (e.g., Aonghusa and Michie, 2020). Alternatively, these predictions could be automatically acted upon, which is how many recommendation algorithms and 'choice engines' function (Johnson, 2021; Sunstein, 2024; Thaler and Tucker, 2013). In the former arrangement, the predictive AI functions as a tool, its output only affecting the world when given motive power by a decision-maker. In the latter instance, the predictive AI functions as a machine, possessive of its own motive power to act on the output via an automated system such as a recommendation banner.

Generative AI utilises machine learning technologies to build models capable of transforming user inputs into informationally greater outputs, perhaps and often in media different to that of the input (Sætra, 2023). The prompt to 'write a short children's story' will often result in several paragraphs of text containing characters, places, and other details absent from the prompt, and thus constituting greater informational complexity.[30] The prompt to 'create a front cover image for my short children's story,' will realise

[29] An added reason for the distinction I will make is that, in recent years, there has been much discussion of how 'AI' can be used to fulfil this or that task, solve this is that problem, and be integrated into this is that domain. These discussions have, as above, been inspired by progress in generative AI capabilities. This is fine. But in most areas where 'AI' is being touted as a new wonder technology, predictive AI is already utilised, or if not utilised, could have been several years prior. When one wishes to talk of the potential to use AI for its generative purposes, one should talk of *generative AI*, because *not all AI is generative*.

[30] The notion of generative AI simply being a predictive AI 'run backwards' is reflected in this idea of greater informational complexity. Predictive AI 'strips away' information from training data to determine underlying statistical relationships from which predictions can be made. Reversing this process implies the 'adding back' of information.

an output in a different medium than the input (an image, compared to text). At present, generative AI is often used as a tool, relying significantly on the motive power of users to apply the outputs to affect the world. Increasingly, however, generative AI is becoming more like a machine, with discussions of generative AI advertisements being integrated into online advertising systems (Thakur, 2024), as well as the growing use of generative AI customer service chatbots (Marr, 2024), and misinformation chatbots on social media which also utilise the technology (Marcellino *et al.*, 2023).

From a technical perspective, the distinction between predictive and generative AI may be trivial. Both use neural AI architectures to estimate probability distributions. But from a behavioural perspective, there is an important difference. As above, the sudden ascendency of generative AI (in part) reflects the saliency of generative outputs.[31] Meanwhile, the lower profile of predictive AI shows a more subtle behavioural effect, and thus warrants different lines of inquiry. Though, as one might expect, there is overlap between the two. For instance, if one can predict that someone likes the colour green (predictive AI), one may wish to automatically change a website to prominently feature this colour (generative AI). As such, at times this division of predictive and generative is less useful than articulating the overall behavioural implication of the application of AI technologies.

*

With the above definitions now presented, the rest of this book may begin.

[31] I say 'in part' because another substantial component of recent AI adoption is the development of user interfaces (UI) which make it much easier for ordinary people to use these technologies. See Chapter 4.

Chapter [2]—Personalisation

> "In the society that has come into existence, since the Middle Ages, one can always avoid picking up a pen, but one cannot avoid being described, identified, certified, and handled—like a text. Even in reaching out to become one's own "self," one reaches out for a text."
>
> —Ivan Illich and Barry Sanders, *The Alphabetization of the Popular Mind* (1988, p. x)

The Problem of Heterogeneity

Modern behavioural science is, in a manner of speaking, a social technology.[32] It utilises scientific results concerning decision-making and human behaviour to design social interfaces with

[32] By a social technology, I essentially mean a technology that acts directly on people, groups, communities, and societies, rather than on raw materials like metal, wood, fossil fuels, and so on. In its broadest use, one could argue that something as innocuous as a good argument is a social technology, insofar as that argument is crafted in accordance with rules of rhetoric to achieve an expected (i.e., predicted) outcome.

Some might object to my use of the term 'social technology' in relation to modern behavioural science. Perhaps, so the argument might go, it is far too mechanistic and discounts the very legitimate efforts by some to *not* treat people instrumentally, or as simple objects to be moved around a maze (e.g., Sunstein, 2017). My retort here is twofold.

Firstly, as I will argue throughout the remainder of this book, once the technological possibilities of AI are integrated into behavioural science, one *must* develop an understanding of behavioural science as a social technology, less one fail to grasp the true nature of the beast. My argument is thus that, while 'analogue' variants of modern behavioural science may be able to stake this defence, 'digital' variants, via scale arising from digitalisation, render this defence tenuous. See Chapter 5. I am also sceptical of the defence one might stake given some uses (and misuses) of language, as explored somewhat in Chapter 4.

Secondly, as I will touch on only here and there throughout this book, modern behavioural science has interesting philosophical and historical connections to techno-social fields such as cybernetics (for writing on the techno-social perspective of modern behavioural science, see Frischmann and Selinger, 2018). Both view social systems as entities to be managed, while wrestling with the desires for and the consequences of autonomy within said systems. Both developed from research closely linked to post-war RAND Corporation work, and as outlined in Chapter 0, have links to programmes such as the early years of AI research. And both, interestingly, deal in the same language. Cybernetics is derived from the Greek for *steersmanship* (Medina, 2014; Wiener, 1950), while modern behavioural science is often described as "steering" individuals towards better choices (Sunstein, 2015, p. 417). Feedback, a key component of cybernetics, is also used to justify nudging and choice architecture (Thaler and Sunstein, 2008). As Davies (2024) argues, cybernetics would likely have retained its close links to social science, rather than being absorbed into AI research and science fiction stories, had the founders of the discipline not been so good at building mechanical versions of the social phenomena they were actually interested in.

predictions as to the behavioural outcomes of these interfaces. One name for these interfaces is *choice architecture* (Thaler and Sunstein, 2008).

Choice architecture is unavoidable (Sunstein, 2017, 2013). For an option to be presented, *some* presentation must be given. Choices cannot exist in the aether, or in some abstract space accessible only through pure thought. Items must be presented on menus, on websites, on picnic tables, and so on. And with such presentation comes the opportunity to influence. Vegetarian items on menus may be isolated from other choices, signalling that they are 'weird' or 'abnormal,' and thus discouraging their selection (Bacon and Krpan, 2018). Choices may feature prominently on a website landing page, surrounded by bright colours and captivating graphics, signalling they are desirable, or at the least, planting themselves in a person's mind as items to be considered, if only to be rejected (Benartzi, 2017). And on the picnic table, those snacks which sit within arm's reach may seem to be more desirable than those that require a slightly greater stretching of the arm (Bauer *et al.*, 2021).

Choice architecture often leads choices to be made through information that is not especially relevant to the choice itself (Thaler and Sunstein, 2003). This may be a generous interpretation. Less generously, one might argue that this is a value-judgement, and an arrogant assertion regarding the fallibility of human decision-making (Gigerenzer, 2015). For instance, choosing something because it is easy is not necessarily a bad strategy to adopt, especially when one is generally indifferent to all options, or when there is some pressing factor which means that a particular choice cannot demand a premium on time, or attention, or both.[33]

[33] These arguments are largely informed by Simon's (1997a) original arguments for bounded rationality. Simon, contrary to modern behavioural science, made few value judgements as to the quality of choice, instead choosing to comment on the *method* of choice, and noting that overall, this method leads to *viable* choices which allow decision-makers to get closer to their objectives. He did not disregard the possibility that people made poor choices, or that bounded rationality may be a factor in poor choices. For instance, he noted that difficult choices in one domain may influence a person's choices in a completely separate domain, possibly in a way that leads to worse choices in both. But—crucially—Simon concerned himself less with individual choices, and more with everyday behaviour and problem solving. Undoubtedly, across a typical day, bounded rationality and 'choice architecture' (even when not consciously, paternalistically designed) helps more than it hinders. The notion of 'poor choices' and the need for choice architectural interventions comes about, in part, from a particular focus on *singular* choices, not everyday behaviour.

Kahneman (2011) would later develop ideas around 'reasonableness' in decision-making as a counter to some criticism of the modern behavioural science perspective on 'errors' or 'irrationality' in decision-making. One might argue 'reasonableness' is, to a large extent, an independent restatement of some of Simon's perspective, several decades after Simon originally articulated it. Also see Sen (2002).

These perspectives on the function of choice architecture thus belie different criticisms of how intelligent behaviour should be encouraged and influenced, insofar as the term is defined in previous chapters.

The purposeful design of choice architecture is a central concern of modern behavioural science, and when such designs are aimed at supporting individual wellbeing, without inhibiting or mandating choice, these designs may be called *nudges* (Thaler and Sunstein, 2008).[34]

Nudges have been quite successful, at least insofar as they have penetrated government and other policymaking circles (Mills and Whittle, 2025). Evidence of the effectiveness of nudges and related behavioural science interventions remains contentious, with a variety of arguments for their respective positions given by both advocates and critics.[35] For instance, nudges have been criticised for producing only small effect sizes, suggesting that even if changes in choice architecture affect *some* behavioural change, it is rarely large enough to achieve substantial policy objectives, say in areas such as ecology and pro-environmental behaviour (Chater and Loewenstein, 2023; Nisa *et al.*, 2020). By contrast, nudges have been promoted given their cost-effectiveness, as choice

[34] For the sake of narrative and accessibility, the term 'nudge' and variants thereof will be used quite broadly throughout this book, though strictly speaking, nudges conform to the criteria of promoting decision-maker wellbeing without restricting freedom of choice—what is known as *libertarian paternalism* (Thaler and Sunstein, 2003). An instance where one might use the term 'nudge' without an example conforming to this narrow definition may be when choice architecture influences a decision-maker to choose harmful options. In such an instance, one might use more exotic terms, such as 'dark nudge' (Newall, 2019), 'dark pattern' (Brignull, 2011), 'sludge' (Thaler, 2018), and so on. Where these terms are useful, such terms will be used in this book, too. Though, often, it will be easier to say 'nudge' with clarification in the footnotes, if *really* warranted.

[35] One interesting meta-analysis disputes the effectiveness of nudging, suggesting that the evidence of effectiveness is an artifact of selective publishing by academic behavioural scientists (Maier *et al.*, 2022). This has been disputed from several perspectives.

One criticises the comparison of different choice architecture designs (Hallsworth, 2022). Just as it would be inappropriate to compare a house and a warehouse, so too is it inappropriate to compare, say, a social norm nudge to a present biased nudge. Indeed, alternative meta-analyses which focus on congruent designs do find positive effects from nudges (Jachimowicz *et al.*, 2019). Furthermore, one study of nudges published by practitioners also finds positive effects, suggesting that the notion that evidence of nudges is simply a product of publication bias is false (Della Vigna and Linos, 2022).

Still, several meta-analyses of nudges and related methods report modest effect sizes, suggesting that even if there is a positive effect of these interventions, they should be understood within the context of the policy outcome they are being used to affect (Mills and Whittle, 2025). This is the crux of another important criticism from another recent, controversial paper (Chater and Loewenstein, 2023). Some elaboration on this point will be offered in Chapter 4.

At the time of writing, the best response to the question, 'do nudges work?' is probably, 'yes, sometimes, depending on your definition of *work*.'

architectural interventions tend to be cheap. Thus, even if effect sizes are small, the effect per cost outlay may be substantial (Sunstein, 2013a).[36]

One criticism, related to the effect sizes of behavioural science interventions, is that of the 'problem of heterogeneity' (Mills, 2022a). Stated simply, because people are different (i.e., *heterogenous*), the same behavioural intervention may have significantly different effects on people. Some people may be especially receptive to the intervention, while others may demonstrate no behaviour change. In some instances, one might imagine that the intervention would induce a behaviour contrary to the desired outcome. In short, one size of nudge does not fit all people, and so one-size-fits-all interventions may be less effective than *personalised* interventions which accommodate the differences—*heterogeneity*—between people.

Various evidence suggests that the problem of heterogeneity is a confounding aspect of modern behavioural science interventions. One interesting study (Thunström *et al.*, 2018) examined the effects of a reminder nudge on the spending habits of shoppers. Shoppers were reminded of the opportunity cost of purchasing a product—in this study, locally-sourced honey. If they purchased the honey, they would no longer have funds to buy other items, and so purchasing the honey would cost them (in addition to the retail price) future purchasing opportunities. The intervention worked. While it had only a slightly positive effect, it reduced spending in a treatment group relative to a control. As a headline finding, such a result suggests nudges could be used to encourage those who wish to save more to spend less.

However, the study *also* recorded the attitudes of shoppers towards spending, allowing shoppers to be grouped into 'spendthrifts' (those who tended to spend too much) and 'tightwads' (those who tended to spend too little). Investigating the effectiveness of the reminder nudge by group, Thunström and colleagues found that the nudge had no significant effect on the spending behaviour of spendthrifts. One explanation for this is that spendthrifts, by virtue of regularly spending too much, have accumulated experience of facing opportunity costs, and thus amassed a kind of 'tolerance' to such costs. By contrast, the tightwads significantly reduced their spending when nudged to do so. Again, the explanation comes from the notion of 'tolerance':

[36] Of course, if the absolute effect size remains too small to realise a necessarily large change, and if effect sizes are not additive, then it is largely irrelevant whether an intervention is cheap or not. A cheap intervention that fails is still a waste of money.

those who typically spent little had little experience of opportunity cost, and so were overly receptive to this message.[37]

Besides being an elegant piece of research, this study highlights two aspects which demonstrate the 'problem' at the heart of the 'problem of heterogeneity.' *Firstly,* the overall *average* effect of a behavioural science intervention may actually hide how the intervention's effect is *distributed* across the sample. In the spending study, while there was a slight overall effect, the reality is that there was a reasonably large effect for some, and a negligible effect for others.[38] *Secondly*, even if one assumes the effect of a nudge is, *on average*, welfare-enhancing, it may actually be *harmful* when the distributional effects of the nudge are considered (Sunstein, 2022a). For instance, in the spending study, those who needed nudging the least (tightwads) were those who were most effected, while those who were most likely to benefit from the nudge (spendthrifts) were least receptive to it.

Further evidence of the problem of heterogeneity can be found in various additional studies,[39] and while a minority of

[37] Little experience of opportunity cost, *insofar* as they lacked the funds to pursue alternative choices. Strictly speaking, the choice *not* to spend could still be said to carry opportunity costs.

[38] One interesting aspect of this discussion is just how rarely an 'average' person can actually be found. One story comes from Gilbert S. Daniels (1952), who was employed by the U.S. Airforce to design jumpsuits and similar equipment for pilots. Daniels and his team gathered large amounts of data on the physicality of pilots, and very soon found that by designing a jumpsuit which conformed to the average of each variable captured (e.g., height, waist-size, etc.), *no one* in their sample would fit comfortably.

This reflected a much older phenomenon attributable to the Belgium statistician Adolphe Quêtelet, who noticed that as a sample was successively reduced so as to retain members who conformed to the average of a given variable (again, say, height, then weight, etc.), very quickly *no one* would be left in the sample. As one study put it, the 'average user' is a "myth" (Egelman and Peer, 2015), though one which has benefits in statistical analysis, and which comes easily as a mental benchmark for many of us. Nevertheless, Daniels (1952, p. 5) cautioned against thinking in terms of averages, and encouraged one to think beyond it: "The tendency to think in terms of the "average man" is a pitfall into which many persons blunder."

This, ironically, is reflected in the full story of Quêtelet's work. Quêtelet was an early *social* statistician who believed that by measuring and comparing the 'average man' of different nations, nations could be ranked and social policies developed. This, of course, is premised not just on the belief that an average person actually exists, but that one could objectively determine, say, that it is better to be taller. Quêtelet's 'blunder' into fetishising the average man of respective nations was simply the first of several blunders which would evolve into the eugenics movement of the late nineteenth and early twentieth centuries (Sposini, 2019). For instance, the eugenicist Francis Galton believed averaging the physical features of convicts could be used to predict whether a person was likely to commit a crime, or not (Salvaggio, 2023). I will return to Daniels later in this book.

[39] For a selection of studies, readers are invited to consider Brown *et al.* (2022); Kim *et al.* (2023); Laffan *et al.* (2024); Mrkva *et al.* (2021); Murakami *et al.*, (2022); and Thunström (2019). This list is not exhaustive.

studies engage in 'heterogeneity analysis' (Beshears and Kosowsky, 2021), interest and attentiveness to the role of heterogeneous effects is growing (e.g., Byran, Tipton and Yeager, 2021; Hecht *et al.,* 2023; Krefeld-Schwalb *et al.*, 2024; Mills and Whittle, 2024a). That the problem of heterogeneity exists, though, does not address the question: *why does the problem of heterogeneity matter?* One might even ask: *why is it a problem at all?*

The immediate answer is that if the problem of heterogeneity could be solved, interventions could be more efficiently deployed, and in such a way as to reduce the potential harms they cause (Sunstein, 2022a). From the perspective of modern behavioural science, this would lead to better individual choices and more equitable social outcomes. Personalisation through predictive AI is a compelling solution to this problem (Mills, Costa and Sunstein, 2023). This will be the focus on this chapter henceforth.

How to Personalise a Nudge

Personalisation could solve the problem of heterogeneity as it allows relevant differences between individuals to be incorporated into the choice architectural design.[40] As Peer and Mills (2024, p. 5) define it, "personalised nudging is designing behavioural interventions such that they manifest to the target of the intervention differently, based on some systematic (non-random and non-arbitrary) individual difference variable(s) that is pre-defined by the choice architect" where the "choice architect" is the behavioural scientist who designs choice architecture.[41]

There are several ways of thinking about personalised choice architecture. One perspective is to distinguish between personalising the content of an intervention, versus personalising the method of intervening itself (Dalecke and Karlsen, 2020; Peer

[40] By 'relevant' differences, I mean factors which have some statistically meaningful relationship with the intervention being administered, and the behaviour change being sought. This should not necessarily imply a value-judgement. There may be various examples of, say, gender being a 'relevant' difference when examined statistically. However, normatively, one might object to gender being included in the analysis—one might contend that gender *ought not* be relevant in that particular domain. This reveals a tricky problem for personalised behavioural science; indeed, for AI in general—to what extent should these methods reproduce existing social biases and discriminations because they are predicates of social phenomena, versus opposing such reproduction because they are socially unacceptable? Chapter 4 considers such questions.

[41] One helpful comment may be to distinguish between what one could call 'positive' and 'negative' personalisation. For a computer scientist, personalisation will typically mean unique to each individual. Yet, for a behavioural scientist, personalisation often means *not* the same for everyone. As such, the computer scientist will typically (implicitly) hold a 'positive' view of personalisation (i.e., personalisation *is* X) whereas the behavioural scientist will typically (implicitly) hold a 'negative' view of personalisation (i.e., personalisation *is not* Y).

et al., 2020). Consider a social norm intervention. People are often susceptible to social norms because we care about what other people think of us, and we want to fit in. One method of personalising this intervention is to personalise the messenger of the nudge—the person who is communicating the message (Dolan *et al.,* 2012). As different people have different idols, or people they admire (or dislike), personalising a social norm message to match these individual differences may elevate the message's impact beyond a message delivered by a one-size-fits-all messenger. This is why strategies such as influencer marketing work—one is more likely to listen to messages coming from those whom one likes and trusts (Moon, 2010).

Alternatively, one might personalise the content of the message itself. One interesting study examined how social norm comparisons influenced water usage. Those who used less than the average amount of water received a different comparison to those who used above the average (Schultz *et al.*, 2016). Prior research has found a 'magnet effect' where telling those who are below average compared to their peers in terms energy consumption causes those individuals to *increase* their consumption to a level comparable to their peers (Schultz et al., 2007).[42] In the water study, this personalisation of the social comparison caused those whose usage was above average to decrease it, without introducing a magnet effect in those whose usage was already below average.

In both instances (the messenger and the social comparison), the *content* of the intervention is personalised. Yet, what if some variable suggests that person A is highly susceptible to social comparison, but person B is not? In such an instance, personalising the *method* of nudging may be preferable—say A receives a social norm message while B receives a default option message.[43] One study examining password creation found that stronger passwords were created when different types of nudges were used on individuals with different decision-making styles (Peer *et al.*, 2020). This should be reasonably intuitive—those who are less patient may respond differently to a 'helpful' password tip compared to, say, those who are more risk averse. The former would perhaps prefer strong passwords being recommended to them, while the

[42] A slightly larger magnet effect is found for those who are above average in their consumption, leading to an overall positive effect nevertheless (Schultz *et al.*, 2007).

[43] Relatedly, there has been some discussion of personalising the *audience* to the message, rather than the message to the audience. Say one is going to the cinema. There are perhaps half a dozen movies playing, and each movie will be the same regardless of who is in the audience. Personalisation in this instance would involve predicting which movie each person would enjoy most, rather than *changing the movie* to suit the tastes of each person. Thus, the audience is matched to the content, rather than the content to the audience. See Matz *et al.* (2024) and Matz *et al.* (2017).

latter, by virtue of someone *else* suggesting a password, may be less amenable to this intervention.

As Peer and Mills (2024) argue, personalising the content of an intervention, or the method of intervening, or both, each represent different ways of personalising choice architecture. Additionally, though, they emphasise what one might call the 'material constraints' associated with personalisation. To effectively personalise, one needs to have a) relevant data about people with which to design personalised interventions; and b) a choice environment which is sufficiently *malleable* so as to change based upon what is predicted to be best for that specific individual or group. As Thaler (2021) notes, it is essentially irrelevant saying, 'X is better than Y,' if one can only practically do Y, either because of data or malleability limitations.

With these constraints in mind, Peer and Mills argue that personalisation is not a binary on/off feature of choice architecture, and that while a nudge may be impersonal or not, *there are many ways of personalising*. They outline five 'levels' of personalisation, which form a broad taxonomy of personalised nudges.

Firstly, there is the *named* nudge, which uses trivial information, such as one's name, title, or gender, to appeal to one's sense of self. 'Trivial' here is used descriptively. Peer and Mills argue that named nudges personalise so as to attract attention to the nudge itself. Naming, or appeals to the self, exert an indirect behavioural effect—not a trivial effect *per se*. One example given is Coca-Cola's '*Share a Coke*' campaign, where common first-names replaced standard Coca-Cola labelling. These names brought attention to the product, which then encouraged consumption of the product, while the product itself never changed.

Secondly, there is the *individualised* nudge, which uses non-trivial information, such as one's energy usage data or other consumption behaviour, to personalise the content of an intervention. This is archetypical of the personalised norm discussed above—information is used to personalise the comparison, though that a comparison is made does not change. Unlike the named nudge, the information used in the individualised nudge directly acts on the nudge itself and thus exerts a direct behavioural effect.

Thirdly, there is the *tailored* nudge, where behavioural data (such as one's risk preferences) and contextual data (such as temporal receptiveness) are used to personalise the framing and the context of the intervention. For instance, an intervention around buying an electric car might emphasise the savings of electric vehicle for some, and the costs of combustion engine vehicles for others, depending on their susceptibility to gains or losses. A night-owl might receive an intervention late at night, while an early bird receives it when they wake up. And so on.

Fourthly, there is *targeted* nudging. Within the taxonomy, targeted nudging uses psychometric and decision-making data to personalise the behavioural mechanism used to influence. Some people might care about social acceptance, while others might care more about avoiding losses, and so on. This information is then used to select wholly different nudges to influence behaviour. Yet, targeted nudging differs from prior levels in more subtle ways, too. In a first instance, it makes more sense to talk of *nudging* rather than of a *nudge*, singular. The selection of different nudge mechanisms is a *process* to target a nudge. But that the nudge is targeted does not necessarily reflect in the nudge's design (though it may be named, individualised, or tailored, too). In a second instance, that targeted nudging is about selecting the best *mechanism*, it becomes feasible to imagine that targeted nudging might, sometimes, suggest *no* nudge should be used. Within the set of possible mechanisms, not nudging is viable, and for those individuals who might respond poorly to an intervention, not intervening may—counterintuitively—be the most effective strategy.[44]

Finally, and perhaps most interestingly in the context of AI systems (see Chapter 5), is *adaptive* nudging. Adaptive nudging is targeted nudging with the inclusion of feedback and automatic reconfiguration. Feedback is taken to be the behaviour of those who are nudged—did a person follow the nudge, or not? Alternatively—did a person choose A, B, or C? These feedback data then update what is predicted to influence behaviour, causing an adaption in the personalisation process. Many recommendation algorithms follow this broad principle—if a person consumes a piece of media associated with *x*, increase the likelihood of recommending other media associated with *x*; else, decrease it. Insofar as one is concerned with the question, *did the person do as they were nudged to do*, one might describe such feedback data as *adherence* data.

Peer and Mills argue that each level of personalisation utilises more data, and demands more malleability from the choice environment, to be implemented.[45] This allows the design of

[44] Additionally, Peer and Mills argue that impersonal nudges should be understood as an implicitly sixth (or perhaps zeroth) level of personalisation. As I will return to in Chapter 5, personalisation itself has costs (economic, social, and ethical) which might erode any benefits of personalisation. There may be instances where, despite seeming heterogeneous effects, an impersonal nudge is preferable. See, for instance, Arulsamy and Delaney (2022).

[45] Peer and Mills (2024) do *not* argue that one level is necessarily better than another. Neither do they argue that any or all of these levels are necessarily ethically acceptable or desirable. Their design framework is descriptive, not

personalised nudges to be understood algorithmically. Because data and malleability determine both the breadth of personalisation *within* a level, and the levels available to a choice architect, one can quickly reduce the set of possible designs through if/then criteria and systematically explore those designs which remain. This, in turn, could lend itself to a predictive AI system. For instance, social media platforms are widely known to engage in constant A/B testing of different platform designs and features to determine 'what works' (Hagar and Diakopoulos, 2019). Relatedly, recommendation systems use various data about a person's behaviour to rank all content a person could potentially be shown, before populating their social media feed with that predicted to be most amenable to them (Khambatta *et al.*, 2023; Zuboff, 2019).

Much of the practical benefit of using predictive AI within personalised behavioural science comes from the methodological opportunities the technology unlocks, as well as some of the methodological challenges it overcomes. The 'Big Five' personality scale consists, unsurprisingly, of five dimensions which are said to measure human personality. These are readily recalled with the acronym OCEAN—*openness, conscientiousness, extraversion, agreeableness, neuroticism*. Depending on the specific version used, each person is scored on a one through seven scale for each dimension. These data might be stored, say, as a Python dictionary: {"O": 3.5, "C": 4, "A": 2.5, "E": 6, "N": 5}. Great—*why does this matter?*

A fundamental question in personalisation is: *how* are people different? One approach might be to group people in terms of their extraversion. Extraverts receive nudge design A, introverts receive B. This high/low approach is simple and has been used in various early personalisation studies (e.g., Moon, 2002). Unfortunately, it has a major drawback—how does one determine the cut-off between high (those who receive design A) and low (those who receive design B)? Often, this will be an arbitrary decision.[46]

normative. For instance, if the social costs of accessing some data are greater than the welfare benefits of a 'better' personalised nudge (in terms of potency—see Chapter 5), then a 'higher' level of personalisation is not actually worthwhile. Peer and Mills (2024) instead offer the analogy of a map—just as one island is no better than another by virtue of being further away, one level is not necessarily better than another by virtue of requiring more data, malleability, and technology.

[46] One way would be to use the mean or median. These are intuitive and easy to calculate, but are sample specific, and even if they were not, there is no reason in principle that these values should matter. They just *feel* reasonable. Using the midpoint of the scale has the advantage of being independent from the sample, but again, why should a four on a seven-point scale carry any more, or any less, weight, than any other value? All are examples of the so-called 'pick-a-points' method (Hayes, 2017), whereby potentially reasonable (but still essentially arbitrary) values are used to stratify the sample.

Another approach is to use more advanced statistical techniques, such as moderation analysis (Hayes, 2017). Simply, this allows one to determine the values of a personality scale where a nudge is having a significant effect, and where it is not. Through moderation analysis, one might discover that nudge A is effective only for individuals with an extraversion value of five—this, then, would be a data-driven way of determining who receives A, and who receives B (Peer *et al.*, 2020). Naturally, there is a problem with this approach, too.

Say one's sample contains two people, each of whom have an extraversion score of five, meaning they are both 'extraverted,' in this example. *Are these people the same?* Intuitively, we know they are not—no two people are identical. And, probabilistically, they are not identical, either. While they may have the same extraversion score, it is extremely unlikely that they will have the same score for the other four dimensions of the Big Five.[47] By only focusing on one dimension—which may itself have been arbitrarily chosen—one will readily come to classify different people as being similar, and similar people as being *less different* than they perhaps actually are.[48]

Ideally, a behavioural scientist designing a personalised nudge would like to collect OCEAN data from a large group of individuals. They would also like to nudge this group with a variety

[47] For the sake of discussion, one should assume that these other differences matter within the analysis. It is reasonable to ignore differences between people when those differences lack meaningful explanatory power within the analysis—when they are not *relevant*, as discussed above.

[48] There is something of a rabbit hole here which one could stumble down. For instance, one could ask: 'why use the Big Five, rather than some other personality scale?' One could also ask: 'why (just) measure personality?' This is to say, there is no escaping the decisions made by researchers, experimenters, statisticians, and so on. Eventually, *someone decides* that we are doing apples rather than oranges. While using more data and more sophisticated methods can obscure such decisions (Gitelman, 2013), they remain. The inescapable politics of algorithms will be discussed in Chapter 4 as it applies to algorithms in decision-making. Still, there is more to say.

One might describe moderation analysis within personalised nudging as 'fishing' insofar as one collects a whole bunch of individual level data and investigates what does (and does not) significantly moderate the intervention. The great danger of such fishing—indeed, the great danger of 'the theory of personalised nudging'—is that the possibility of personalisation is never refuted. One difference may not matter, *but no matter!* One can always find some other difference to measure, and to investigate for its relevance to a nudge design. One can always cast another line, so to speak.

It is for this reason that one must take Hayes' (2017) warning seriously. Hayes argues that, when using moderation analysis, one must always support their analysis with *prior* theory. There should always be a reason why one thinks a difference should matter. But more so, Hayes argues (and I agree) that one should reject *every* result that contradicts prior expectations, not just those that do not work, statistically speaking. Otherwise, one is essentially just fishing. This is relevant to predictive AI, too, as this technology essentially allows massive amounts of 'fishing' to occur at once.

of nudges, having determined some way of measuring the effectiveness of the nudge—for the purposes of discussion, assume this will be a simple yes/no variable indicating that the person chose whatever outcome they were being nudged towards (yes), or did not (no).[49] From this setup, the ideal outcome would be to design a statistical model which takes as inputs all OCEAN data (and any other relevant data, such as demographic data), and produces as an output an estimate of how likely each possible nudge design is to positively affect behaviour. Such a model can be created using supervised machine learning, the workhorse of most predictive AI models.

Supervised machine learning works by training an algorithm to predict an outcome from a set of inputs, before adjusting the model's parameters based on whether the model has predicted correctly or not (Russell and Norvig, 2020).[50] Through many iterations of this process, the parameters can be tuned to accurately predict outcomes for examples it has not previously been trained on, producing—in effect—a predictive algorithm, or AI.[51] The utility of such models, from the perspective of behavioural science and personalisation, is that the number of inputs (the input vector) and outputs (the output vector) can be arbitrarily large, and all input data are utilised within the model.[52] Thus, many variables, carrying many instances of individual difference, can be utilised, avoiding methodological challenges such as arbitrarily determining high/low values, or focusing on only one dimension of heterogeneity at a time.

It is easy to be enthusiastic about this perspective, though there are various points of caution. For instance, despite the technology to test such approaches having existed for several years (at the time of writing), testing of these methods remains scarce in behavioural science, likely because of limited skills and technical challenges given current laboratory resources (Mills, Costa and Sunstein, 2023). Examples of where such approaches have been

[49] Other measures could be used. For instance, in the password setting example, it could be some measure of password strength. In the water usage example, it could be a measure in the change of water consumption. The above described aligns with the idea of *potency*, discussed more completely in Chapter 5.

[50] By parameters, I mean the model's weights and biases. For the sake of discussion, it is helpful to think of neural AI models, like supervised machine learning, as weighted average sums. If $X = w_1A + w_2B + w_3C + Z$, where A, B, and C are inputs, then w_1, w_2, and w_3 are weights. The model's bias is a value which might be introduced into a model to encourage it to have some particular behaviour or tendency (hence *bias*). In this case, this is the value Z. This is a simplification, but a helpful one for the purposes of this book.

[51] The training, tuning, and then testing on new observations is why this approach is called 'supervised' learning—the model is told what the right and wrong answers are by a person.

[52] Hence why, above, I showed how the OCEAN data might be formatted for use in a machine learning model.

used also raise questions of social efficacy. For instance, supervised machine learning has been used to predict political beliefs (Kosinski, 2021) and sexual orientation (Wang and Kosinski, 2018), amongst other traits (Kosinski *et al.*, 2013; Youyou *et al.*, 2015). These studies have garnered some controversy, in part because of the potentially disconcerting implications if one takes them seriously (e.g., that an image can reliably predict something personally and politically sensitive, such as one's sexuality), and in part because they demonstrate a shallow link between theory and rapidly developing methodology. This latter point should be emphasised.

While supervised machine learning offers advantages insofar as it readily increases the amount of data which can be built into predictive AI models, this in itself may *disincentivise* proper data collection and experimental practices. On the one hand, formulating theory and testing various hypotheses can be tedious and time consuming, but through that process, one develops knowledge which could have applications in other areas. On the other hand, collecting as much data as possible (big data) and throwing it all into one model will certainly increase the likelihood that one finds some statistically compelling relationship between inputs and outputs (as it is very likely that the main causal variable is in the dataset), but through this process, one may struggle (or consider it unnecessary) to identify *precisely* which variables matter, and which do not.[53] Such a scenario has prompted some to question whether science is entering a post-theory era, where all that really matters is data (Anderson, 2008).

Yet, it is probably unwise to write-off theory. Firstly, the collection of data has social implications, particularly when considering individual-level data to measure individual differences. People value privacy and should generally be respected. Even where the economic costs of collecting data are low, owing to the proliferation of various digital platforms and applications, the social costs of data collection should not be ignored (Frischmann and Selinger, 2018). Secondly, while the costs of *collecting* data are relatively low, the costs of *using* data can be quite high, both economically for businesses and governments, and environmentally (and again, socially) for communities which must suffer the effects of large data centres (Hogan, 2015; Lehuedé, 2024). Finally, the insights behavioural science produces are, broadly, designed to affect human behaviour and shape the lives of individuals. As citizens, if not as scientists, behavioural scientists should respect the rights of their fellow citizens to understand why

[53] Lanier (2011, p. 98, original emphasis) describes this as "big *n* as a substitute for judgement."

and how they are being influenced, what the implications of such influence are, and to whose benefit.[54]

In each instance, theory can help inform policy and practitioner decision-making about how, when, and why to use predictive AI models to design personalised nudges. The methodological advantages of AI within behavioural science should not be used as an excuse to sideline theoretical perspectives which—in more instances than is often appreciated—render useful the 'practical' insights of empirical work.

It is interesting to note that the combination of predictive AI with personalisation of choice architecture is yet to be widely implemented. *Yet.* As of writing, while we might occasionally be subject to A/B testing or experimental features on websites, social media platforms and other websites or online services are not currently 'morphing' to suit our individual preferences in terms of colour, layout, text size, text density, and so on.[55]

Nevertheless, evidence suggests such personalisation could be introduced (Bucher and Dayan, 2023; Hagar and Diakopoulosu, 2019; Hauser *et al.*, 2014; Hauser *et al.*, 2009; Liberali and Ferecatu, 2022; Reinecke and Gajos, 2014).[56] In the past, various barriers to

[54] All such 'social costs' should be factored into any debate about the use and effectiveness of personalised behavioural science interventions. If it is true that some degree of personalisation leads to a greater effect size, and thus behaviour change at the population level, one falls into a trap of blinkered thinking if one decides this necessarily justifies the use of the intervention. In simple terms, if the additional costs of personalisation outweigh the benefits personalisation is anticipated to bring, then impersonal interventions would be preferable. This is explored further in Chapter 5.

[55] For instance, Hagar and Diakopoulos (2019) discuss how online news headlines could be automatically changed to optimise for whatever outcome a news publisher wishes—say, click-through-rate on a social media website. Whether such an approach would use positive personalisation—with each person seeing a different headline—or just negative personalisation—with the headline simply being that which the largest group are most likely to click on—remains to be seen.

An example of each, analogous to the news headline idea, can already be seen. YouTube offers creators the opportunity to change the thumbnail of their videos at any time, and creators have been known to experiment with different thumbnails following the release of a video, in response to real-time viewership data. This is negative personalisation, as the thumbnail is the same for everyone, though presumably, the content creator is searching for the one which appeals to the most viewers. Similarly, Netflix changes the preview image of its content based on what it predicts will be most attractive to individuals. This could be described as positive personalisation, at least insofar as two Netflix users may, at any given time, see two different content previews (though it is unlikely each preview is unique to each user).

[56] Substantial personalisation occurs in the selection of content for us to see on websites, from social media to shopping websites, and—famously—in the advertisements shown to us (Zuboff, 2019). But, as Peer and Mills (2024) argue, recommendation algorithms, while they personalise, are not really personalised *nudges*. *What* is shown changes, but *how* it is shown does not. The penultimate section picks up on this discussion.

such morphing existed which demanded technical possibility succumb to sticky reality. Increasingly, though, barriers such as access to data are being erased (Zuboff, 2019), while advances in predictive AI do not just mean that the computational barrier to data analysis and automaticity is being overcome, but some aspects of the skill barrier to using these technologies are falling, too. The emergence of personalised behavioural science, powered by predictive AI, as the *default* approach within behavioural science, is a real possibility (Mills, Costa and Sunstein, 2023).

Generative Design

A brief comment is warranted regarding generative AI. It is worthwhile to consider that generative AI could also be used within the personalisation of nudges. Individual-level data could be provided to a generative AI system which is charged with generating bespoke website designs to influence people. Alternatively, a predictive AI might predict design techniques to use (e.g., bright colours, simple text, gain framed language, etc.) which is then given to a generative AI system as a prompt to generate corresponding website designs. As people navigate these websites, websites could be redesigned in real-time in response to the behaviour of users. This collective process could be called *generative design*.

At the time of writing, generative design has not been developed. Or, if it has, has not been widely adopted or, indeed, promoted.[57] One example to highlight is that of *website morphing*. Studies into website morphing use various demographic and cognitive details about people to automatically adapt the layout and content of websites to appeal to users, or to encourage users to engage with websites in particular ways (Hauser *et al.*, 2014; Hauser *et al.*, 2009; Liberali and Ferecatu, 2022; Reinecke and Gajos, 2014). Such a technique is not the same as generative design. Rather than a generative AI developing designs in real-time to influence people, website morphing assembles bespoke designs from catalogues of web design assets in a statistical manner. This can be onerous, whereas the likely ambition of a generative design approach is the *automaticity* of the design-redesign process.

Nevertheless, besides the challenge of *demonstrating* generative design, another challenge remains. This is the challenge of consistency and reliability. The concept of malleability can be useful to explain this challenge. In some ways, generative AI is a

[57] O'Toole (2024) has speculated on the possibility of websites which use generative design and the emergence of a more 'adaptive' internet. They envision websites which change in response to user behaviour, primarily through interactions which AI chatbots. Initial prototypes suggest further development is needed to realise effective, real-time generative design. But that such ideas are being considered is noteworthy.

tool which offers a behavioural scientist tremendous malleability. Images, sounds, and videos can be generated simply through the text prompting of a choice architect. Most outputs can, in principle, be generated, offering an almost universal paintbrush for behavioural scientists to play with.

Yet, the appearance of malleability does not overcome the current challenges to achieving said malleability. Glass, after all, is a liquid, but one still cannot swim in it. Generative AI, while it may in principle be able to generate bespoke visuals for a personalised nudge or an entire personalised website, is often too *inconsistent* and *unreliable* to be a viable tool.[58] Undoubtedly, there will be some features of a website which *must* remain, or which one would not wish to go without or see compromised. There is also the risk that generative design breaks a website, denies a person website functionality, or otherwise confuses the look and flow of a website to the point that it is inhibitive. Generative AI, like telling a child to 'go wild', reveals that in some instances, too much malleability undermines any structure or consistency, to the detriment of the whole endeavour. To this end, the downside of website morphing—substantial investment in building the architecture to enable successful morphing—is also an upside when one considers the likely fragility of generative design approaches.

Undoubtedly, efforts will emerge in this space. It does not seem unreasonable to speculate that some constraints could be placed on a generative design approach to make it more consistent and reliable. Yet, for now, the role of generative AI in personalised behavioural science must be diminished, at least so the more important role of predictive AI can be elevated. This is a space to be watched.[59]

Personalised Paternalism and Selection Bias

The previous section focuses on personalising the content of behavioural science interventions, and on the method of intervening itself. Both are broadly *design* aspects of the intervention. However, there is a third element of an intervention which could be personalised—the intended *outcome* of the intervention (Mills, 2022a).

[58] One tool I have played with suffered twofold from designs which failed to ever manifest as a coherent whole—something which is quite jarring given various sophisticated websites available today—and the speed at which websites adapted to my actions and prompts—often, ten to fifteen seconds would elapse before a page would load, reflecting the substantial computational burden attached to generative design.

[59] The topic of generative design will return, though somewhat indirectly, in Chapter 5 and the discussion of *autonomous choice architecture*, to which generative design could be a compelling example.

For instance, say an intervention is being used to default people into choosing fruit rather than an unhealthy dessert. This is to say, people will receive fruit unless they opt-out and choose an unhealthy dessert instead. But some people will prefer apples, and others will prefer oranges. Someone who prefers oranges may opt-out not because they want the dessert *per se*, but because they *do not* want to receive an apple. This is, once again, the problem of heterogeneity. As such, one solution could be to personalise the outcome to which an intervention encourages—in this instance, setting oranges as the default for those who prefer oranges, and apples as the default for those who like apples.[60]

Mills (2022a, p. 150) has called this "choice personalisation." Yet, this name masks what is really happening. Sunstein's (2013b, p. 1871) prior idea of "personalised paternalism" is a much more accurate description of this aspect of personalised behavioural science.

When one personalises the design of a behavioural science intervention (using any of the 'levels' of personalisation described above), one is making no judgement about *what* a person should receive or should do. There may be some prior judgement about *whether* a behavioural intervention should be used, and to what end. Yet, when personalisation is only applied to the design of the intervention, these matters do not fall within the domain of personalisation. Thus, one can—in some ways—set aside questions of whether the intervention is 'good' or 'bad' and recognise that the purpose of the design is to encourage the outcome which has been determined to be worthwhile to encourage.[61]

[60] As with anything, there is messiness and nuance here. Say, for instance, one is encouraging someone to save more for retirement. One person might be nudged to save 4%, while another might be nudged to save 5%, the difference deriving from some relevant factors about each person. What kind of personalisation is this? On the one hand, the outcome is being personalised—each person is nudged to adopt a different savings rate. On the other hand, only the content is personalised—each person is still being nudged to save. In my work with Eyal Peer, we have often had to explore this question.

If I were to speculate, I would suggest the following. For each decision, there is a range of tolerable outcomes, the rest being intolerable. Insofar as everyone is assumed to have the same range of tolerance, it is the content, rather than the outcome, which is personalised. Insofar as people are taken to have different sets of tolerable outcomes, it is the outcome which is personalised. For instance, let us continue the saving example. For many people, the difference between 4% and 5% is negligible, thus the personalisation here is likely to be content personalisation. If, by contrast, one person was nudged to save 4%, and another to save 40%, this is likely to be outcome personalisation, as for most (though not necessarily all) people, 40% is an intolerable savings rate.

[61] I say, "in some ways" because there is a valid argument to make that one cannot effectively personalise an intervention if one does not know whether the intervention is helping or hindering someone. Indeed, the whole problem of heterogeneity to which the personalised intervention responds emerges from

When one uses 'choice personalisation' or, more appropriately, 'personalised paternalism,' one is engaged in an exercise of trying to predict which *outcomes* are best for someone rather than which *designs* someone is most likely to adhere to. There are good reasons to think this would be worthwhile, from a behavioural science perspective. Firstly, if one is concerned with the effectiveness of an intervention—measured in terms of how many people do as they are nudged—then nudging people towards options they themselves would prefer is going to be desirable. Secondly, and quite obviously, encouraging people to engage in behaviours which are genuinely better for them (*as judged by themselves*; Thaler and Sunstein, 2008) is better than promoting via an intervention some less than desirable outcome.

Broadly, insofar as one accepts the arguments for paternalism (e.g., Sunstein, 2013a), then the use of personalisation within a paternalist programme should warrant little objection.[62] Indeed, personalisation may overcome one criticism of paternalism, namely, the goodness of fit to individuals. If one must nudge everyone towards the same outcome, one inevitably must adopt a utilitarian logic of nudging towards the outcome likely to deliver the most benefit to the most people (Sætra, 2019). Yet, if one can personalise the outcomes to which people are encouraged, then the implicit demand that outliers should compromise for the benefit of the group may vanish.[63]

considering the welfare implications of intervening in a given way. For instance, 'adaptive' nudging requires data about a person's previous response to the intervention to personalise future interventions. This could feasibly include some measure of welfare or utility.

Yet, as discussed in Chapter 5, a simpler measure of adherence is often available—one which does not raise the spectre of whether an intervention is 'good' or 'bad' for an individual. This measure is potency; simply, did a person do as they were nudged, or not? If one adopts a potency perspective, there is little need to actually measure whether the outcome a person receives—whether encouraged or discouraged—actually made their life better.

[62] Or, at least, relatively few novel objections.

[63] There are some caveats here. *Firstly*, this is assuming one is actually able to predict preferences accurately. If not, there is a real chance that not even the utilitarian 'local optimum' is reached, never mind the 'global optimum' of personalised paternalism. *Secondly*, as above, this argument says nothing of the costs of personalisation, which may include greater surveillance and a loss of privacy, as well as demands to be integrated within a technological system one would rather avoid. Indeed, one argument might be that personalised paternalism encourages paternalists to ignore those whose preferences cannot be predicted, perhaps because there is too little data gathered about them, or because they are not technologically involved enough. Thus, some may greatly benefit, and others greatly suffer, from a personalised paternalism programme. *Thirdly*, depending on the predictive model used, outliers may not be as well-represented as the theoretical promise suggests. Indeed, many large language models (LLMs) appear to be biased towards the average of the dataset (Peterson, 2024). This exposes a fourth caveat. *Fourthly*, these promises may not be realisable with the technologies which are currently available to behavioural scientists.

It is no surprise that the explosion in AI technology has been met by a relative explosion in the discussion of personalised paternalism systems within behavioural science. Just as predictive AI can be used to predict which aspects of choice architecture are optimal to personalise the *design* of a nudge, AI can conceivably be used to predict the optimal *outcomes* or *choices* towards which a person should be nudged.

Within behavioural science, such systems are called "choice engines" (Johnson, 2021, para. 1; Thaler and Tucker, 2013, p. 44; Sunstein, 2024, p. 2).[64] These systems fall within the same category of behavioural technologies as recommendation algorithms and targeted advertising systems. The essential difference seems to be the paternalist criterion. Choice engines should encourage people to choose options that will improve their wellbeing. By contrast, recommendation algorithms can have more opaque functions. For instance, YouTube does not recommend people videos to watch because they want to be paternalistic, but because they want visitors to their website to *keep watching*. Recommending videos which visitors are interested in *just happens* to align with this goal. At best, this is dubiously paternalistic. As such, it seems reasonable to say that all choice engines are recommendation algorithms, but not all recommendation algorithms are choice engines.

Certainly, whether paternalist or not, choice engines or recommendation algorithms, powered by predictive AI technologies, are increasingly essential for navigating many services in the online world.[65] This is because the online world is increasingly flooded with information. Social media sites are awash with new posts; media outlets publish dozens of articles daily, which can stick around within the information ecosystem much longer than newspapers ever could; streaming services place immeasurable amounts of content at a person's fingertips, daring them to jump in. The advent of generative AI, for good or ill, is likely to make this information problem only worse, as the immediate cost of producing 'new' information falls dramatically.[66]

In such an environment, tools to help people navigate towards preferable outcomes find themselves in greater demand

[64] Sunstein (2024, p. 1) uses the term "paternalistic AI," though this is perhaps objectionable. It is not the AI system that is acting paternally—the behavioural scientist behind the AI system is using the system to achieve paternal goals, and to act in a paternal way. The AI lacks any morality or responsible; it is an unnecessary anthropomorphism to suggest that the AI system 'itself' is paternalistic. See Chapter 5.

[65] And targeted advertising increasingly a feature of the online world.

[66] This is not to suggest that AI-generated content is cheap in terms of economic or social costs. Economically, many costs are—at present—being absorbed by AI start-ups seeking to build a market position. Socially, individuals are in a constant battle with AI companies over data rights, while environmental externalities push the true costs of AI into the atmosphere.

(Brynjolfsson *et al.*, 2024). Indeed, there may even be benefits to *automating* choices when said choices can be reliably predicted, and where the effort of making those choices is significant. In this instance, having a personalised choice engine or a recommendation algorithm automatically choose some outcome could be beneficial for an individual (Sunstein, 2024).[67]

Yet, central to any benefits of choice engines or recommendation algorithms is the question: *do they actually work*? Different scenarios have different definitions of 'work.' For a recommendation algorithm or a targeted advertisement, the 'work' may be defined as encouraging a person to buy a product which they otherwise would not have. For a choice engine, it may be defined as encouraging a person to buy a product they otherwise would not have bought, which is better for them. Immediately, one can see that choice engines have a higher standard for success than the wider category of 'outcome predicting algorithms.' As such, let us set the choice engine aside for a time, and consider other types of predictive, personalised algorithms, namely, recommendation algorithms and targeted advertising.

Recommendation algorithms predict the post, video, or product to populate in a box on a website. Targeted advertising algorithms predict which advertisement to populate in a pre-specified advertising location on a website or in an app. Both often use predictive AI systems to match a piece of content to an individual based on that person's behaviours, tastes, and interests.

According to Zuboff (2019), targeted advertising began life as a reasonably benign entity. Pioneered by Google, targeting was seen as a way of maintaining Google's reputation for high quality search. If one *had* to see an advertisement in their search, Google hypothesised it would be better for a person to see an advertisement which was relevant to their search. A wholly unrelated advertisement could have undermined Google's reputation for high-quality search results, and thus harmed the company.[68]

[67] Such advocacy often overlooks the important question of: *why is there so much information to begin with?* Those who advocate the use of algorithms and AI systems to navigate information implicitly assume the information being navigated is worthwhile or in some way necessary—if it were not, it need not exist and could simply be eliminated. This idea is considered in Chapter 4. Doing so offers an alternative perspective on the necessity of choice engines and recommendation algorithms.

[68] As another example of how recommendation algorithms and targeted advertising are highly congruent; if one has to see an advertisement, an argument can be made—from a paternalist perspective—that one should be shown an advertisement one might actually be interested in. If one accepts that, sometimes, advertising has a welfare-enhancing effect, then targeting ads *might* have some

While Zuboff's (2019) account should not be rejected out-of-hand, a critical mind will also note the tremendous advantage that the 'narrative' of targeted advertisements could create for a company like Google, and later Facebook. Marketers have, for decades, sought ways to communicate directly to their target market. When one advertises on a billboard, or on the television, one reaches many eyes, but most will look away, leaving much of the money spent on advertising to be wasted.[69] Likewise, when a behavioural scientist working for a government wishes to nudge a group of people to change their behaviour, resources are wasted if messaging must go to all people, including those who *absolutely* will not change their behaviour. This is another perspective which points to the problem of heterogeneity.

The promise of targeted advertising, just as the promise of personalised nudging and personalised paternalism, is a) that the same outcome can be achieved with fewer resources; and/or b) that better outcomes can be achieved with the same resources.[70] Thus, while Google's development of targeted advertising may have been motivated by a genuine desire to maintain the quality of their search product (Zuboff, 2019), the targeting system they subsequently developed also had a compelling—one might even say *enthralling*—narrative to attract would-be advertisers built around the promise of efficiency.

Empirical testing of this narrative is more complicated. A 2012 study found that targeted advertising on social media had a significantly positive effect for advertised brands. Yet, as the authors of the study acknowledge, this effect did not account for 'selection bias' (Farahat and Bailey, 2012). In this context, selection bias must be understood as the percentage of people who would have had a positive response to the advertised brand, *regardless* of the advertisement. This is to say, they have already *selected* themselves to be in the group of people who like the brand.[71] A 2015 study investigated the role of selection bias more deeply. Drawing on a natural experiment at eBay, Blake, Nosko and Tadelis (2015) found that a pause in targeted advertising by the company resulted in no significant change in the number of visitors to the company's website. For the authors, the explanation for this is precisely selection bias—those who were most likely to see an

socially beneficial outcomes, which is the same broad justification for recommendation algorithms. See Brynjolfsson *et al.* (2024).

[69] One, somewhat unsavoury analogy, is that non-targeted advertising is like cluster bombing. It is highly destructive, and resource intensive, compared to the desired, eventual, outcome.

[70] To continue the unsavoury analogy, the narrative around targeting, personalisation, and so on, is like that of a precision-guided missile. It conserves resources and minimises collateral damage while achieving the desired outcome.

[71] This is like nudging someone who would already exhibited the desired behaviour.

advertisement for eBay were *already* those who were likely to visit eBay (Frederik and Martijn, 2019).[72]

To articulate this point, consider the following hypothetical (though hardly unrealistic) policy scenario (Peer and Mills, 2024; Reñosa *et al.*, 2021). A behavioural scientist is tasked with increasing vaccination rates using a nudge. A nudge is designed and is the vanguard of a nationwide pro-vaccination campaign. After a few weeks, around 70% of the population is vaccinated. The behavioural scientist now considers an important question—how effective was this intervention?

On the one hand, it was highly effective—70% of people were now vaccinated. On the other hand, 30% of people were not. In fact, 30% of people were very resistant to vaccination, a small percentage vocally so. The behavioural scientist realises that, in a perfect world, this group should have never been nudged, because they were never going to get vaccinated—in fact, the nudge might have even encouraged some to double-down on their anti-vaccine position (Attwell and Freeman, 2015).[73] Resources and social cohesion may have been spared if the nudge had been targeted better.

Regardless, 70% is still regarded as quite an impressive figure. The behavioural scientist begins typing up the final evaluation report when a thought occurs: *if 30% of people should not have been nudged because they were never going to get vaccinated, what percentage of people were nudged despite the fact they were always going to get vaccinated?* The behavioural scientist, head in hands, deletes the draft of the report. 30% of people had selected themselves *out* of being vaccinated, regardless of the nudge; what percentage had selected themselves *into* being vaccinated regardless of the nudge? This is not a trivial problem—if 65% of people would have been vaccinated regardless, the behavioural scientist has potentially wasted significant resources to achieve a marginal effect.[74]

[72] This is not to imply that some targeting cannot be effective in some instances. For instance, a person who wishes to purchase a fridge exhibits a unique trait for a period of time. Call it *fridge-wantingness*. A fridge retailer may particularly value the specificity of this individual, and thus the matter is not so much about nudging them to buy a fridge (as they have already selected to do so), but to buy *a particular brand's* fridge. Nevertheless, any casual user of the internet will also note that one often receives advertisements at times which are temporally inconsistent with one's behaviour. If this individual, having bought a fridge, continue to see ads for fridges for subsequent days or weeks, those ads are confusing a person highly likely to buy a fridge with someone highly *unlikely* to buy a fridge. This would be a waste of an advertisement (and potentially annoying or unnerving to the person being advertised to).

[73] Hence why, as above, within any discussion of personalised nudging, *not* nudging should be considered a viable design choice.

[74] Again, this is not considering any 'backfire' effects from nudging those that did not want to be nudged (Attwell and Freeman, 2015). While at a certain point

In both the targeted advertising research, where personalisation is used, and the vaccination scenario, where an impersonal nudge is considered but a personalised nudge could readily be substituted, the problem is the same: without accounting for selection bias, many of the 'efficiency' benefits of personalisation which so attract advertisers and behavioural scientists *disappear*.[75]

For all these systems—targeted advertising, recommendation algorithms, and choice engines—their headline promise is to identify and nudge people towards outcomes they would not have previously considered. Else, what is the advantage of these systems in the first instance? Yet, without accounting for selection bias, there is a risk that these systems simply encourage people towards outcomes they *probably would have already chosen*. This is not an efficient deployment of personalisation, though it may often have the appearance of being.[76]

one may believe they are wallowing in minutiae, the question of personalisation does reiterate an argument made earlier—it is really difficult to know how effective any nudge (or any policy) *actually* is.

[75] In an interview with Frederik and Martijn (2019), Tadelis explains that many advertisers are reluctant to withdraw their online advertising, despite the empirical evidence that the audience being reached is one that advertisers already have. They note how the dominance of online advertising today means *not* advertising on Google or Facebook is a substantial business risk. Thus, while the narrative around targeted advertising may be tainted by the problem of selection bias, it may also be *too good* of a narrative for advertisers to have the confidence to challenge.

[76] Perhaps slightly more formally, the ideal recommendation algorithm wants to identify some outcome x given that some person A likes outcome y. In instances where preferences for x and y are highly correlated, or there is substantial overlap, this approach will not account for selection bias. This is partly why predicting *novel* outcomes given *known* preferences is so tricky—the novel outcome must be similar enough to priors to be an accurate prediction, while being dissimilar enough to be novel.

By contrast, if a recommendation algorithm wishes to identify some outcome x such that some person A is likely to choose it, a reliable way of estimating this is simply to examine how many times previously A has interacted with x. Mathematically, then, $A(x) = \frac{count(x)}{N}$, where N is person $A's$ total interactions over some timespan. This is computationally much easier to calculate, but essentially sidelines the more important question of whether A should be nudged to choose x. If $A(x)$ is very high, A probably should not be nudged—they are already likely to seek x of their own accord. If $A(x)$ is low, A *probably* should not be nudged, either—the low probability might signal a dislike of x. Fundamentally, one needs to land on some meaningful threshold where A is simultaneously likely to enjoy x, but *unlikely* to choose it on their own. The simple calculation of $A(x)$ offers little recourse to this more substantial problem. Indeed, it might obfuscate it. If $A(x)$ is low, it might be because A dislikes x. Alternatively, it might be that x is highly novel to A. Given novelty is what we are seeking, one interpretation of $A(x)$ could simultaneously encourage us to nudge A towards x, and to *not* nudge.

Given this problem, from the perspective of a company such as Google or Facebook, it may be tempting to simply nudge people who are already likely

The alternative—nudging people towards outcomes they have not considered, but also would like had they considered them—is a much trickier proposition. Thus, in the abstract, this is potentially a useful application of predictive AI. At the least, it demonstrates an application of continued AI development, if more sophisticated models can produce predictions which result in better outcomes for individuals.[77]

Returning to choice engines, it was noted above that the challenge they faced is not just to predict outcomes a person would embrace, but to predict outcomes that would leave those people better off. From one perspective, this additional requirement is just another layer of predictive demand to be integrated within a predictive AI model. The past few pages do not need repeating. It may thus be fortuitous to consider an alternative perspective.

Advocates of choice engines set AI technologies the onerous task of predicting welfare-enhancing outcomes by sidelining the human decision-maker within the decision-making process. Long before discussion of choice engines—indeed, long before advances in AI technologies and recommendation systems—Simon (1987a, 1987b) explored the question of how people could collaborate with AI systems to make better choices. Simon characterises AI systems as 'expert support systems.' For Simon (1981), people are often able to make satisficing or 'good enough,' choices. Experts may even make excellent choices.[78] But often, people struggle to exercise their expertise because an excess of unnecessary or distracting information clouds their judgement, perhaps resulting in mistakes, or in choices that people recognise with hindsight were not preferable.[79]

to choose x towards it. If advertisers do not account for selection bias within the results they receive, or worse, are scared to *not* advertise on these platforms and use these algorithms, then the impression of high effectiveness remains, and the conceptually tricky problem of accurately predicting novel outcomes is sidestepped for a much easier proposition.

[77] One must note that much of the empirical work on recommendation algorithms and targeted advertising discussed here is old, relative to advances in AI technologies. There is a real possibility that technological advances have allowed some of the problems discussed in this chapter to be overcome. Nevertheless, given the importance of these challenges when assessing the claims of personalisation, there is merit in highlighting them. Furthermore, for institutions which have yet to develop substantial AI competencies—such as in many governments—these challenges may emerge as novel ones, even with the latest technology.

[78] This is something which Sunstein (2023) acknowledges in his discussion of decision-making algorithms, though without explicitly drawing on Simon's work.

[79] One should acknowledge that Simon (1997a) is quite broad in his perspective of onerous information and distractions. Information might relate to the choices themselves. This would then be a problem of too much (potentially relevant) information. Sunstein (2023) typically emphasises this perspective. But for Simon, information could also relate to choices, activities, events (etc.) which are

Simon (1987a, 1987b) suggests that AI systems could be used to filter out unnecessary information, freeing people to exercise their own judgement in more ideal decisional settings. Furthermore, AI-driven personalisation could be used to learn what information different people do and do not use. Thus, one would not have to pre-determine what information is and is not relevant.[80] One might imagine a decision-maker, say a doctor, uses only three pieces of information when taking a decision, despite being given ten. An AI system, over time, might filter out these seven unused pieces. Then, again over time, the AI system might learn that the doctor occasionally seeks out a fourth piece of information, and thus the system might come to filter only six pieces on these occasions. Doing so says *nothing* about the doctor's decisions, beyond recognising that more convenient decisional settings are likely to help the doctor decide, whatever that decision may be. In essence, filtering information to enable people to choose better options, *as judged by themselves*.

Early work on choice engines could align with Simon's proposal. For instance, Thaler and Tucker (2013) emphasised how financial documents could be personalised through choice engines to make them more accessible to ordinary people. A confusing phrase may intimidate a person, discouraging them from asking for clarity.[81] A simpler expression of the same detail might prompt advice-seeking, leaving the person, on the whole, more informed. Such a proposal does nothing to direct the person in their ultimate decision, but rather uses the choice engine to support the person in reaching a decision themselves.

Choice engines only face the challenge of predicting outcomes when designed to exhibit personalised paternalism. This challenge has no immediate solutions—though, methodologically, advances in AI and behavioural science do not foreclose the possibility of developing such systems. This is an idea that subsequent chapters will return to, as well as the implication therein.

Putting a Pin in Personalisation

This chapter has demonstrated is that personalisation is—*fittingly*—heterogeneous. There are many ways to personalise an intervention. While personalising outcomes *might* be worthwhile in

wholly unrelated to the immediate choice at hand. This perspective may have some links to decisional noise—again, see Sunstein (2023).

[80] Indeed, it is often attempts to engineer information environments from above which can lead to problems of too much information.

[81] It might even be so confusing they do not realise they would benefit from further detail.

some instances, a more expansive view of personalisation within behavioural science is likely essential.

This chapter has specifically considered AI-powered personalisation in behavioural science. It has considered how AI systems—with an emphasis on predictive AI—could be used to promote personalised behavioural interventions. In the chapters to come, an engaged reader might wonder why personalisation has been given an individual chapter. Two reasons are worth emphasising.

Firstly, personalised behavioural science is the vanguard application of AI technologies within behavioural science. As a broad category, it is regarded as a major application with exciting possibilities (Mills, Costa and Sunstein, 2023). That personalisation may solve the problem of heterogeneity and boost the effectiveness of behavioural interventions is a significant motivator of many in the field. It is also a frequent promise of technologists, who paint the picture of a more personal, customisable, and 'frictionless' world attuned to each of our individual preferences (Frischmann and Selinger, 2018). To this end, special attention towards personalisation is warranted.

Secondly, personalisation is exemplar of a wider phenomenon of behavioural technology and the mechanical philosophy embedded in (if not surrounding) behavioural science and the notion of intelligent behaviour. Many of the applications of AI to be discussed—bias identification, AI chatbots, simulations, algorithmic decision-making—draw from or align with a broad programme of data-driven, AI-powered personalisation. There are few perspectives on AI technologies within behavioural science which do not draw upon some of the precedents set by the topic of personalisation, from the narrative of 'greater efficiency' to the imperative for more intimate data.

Chapter 5 will return to the question of personalisation more explicitly, focusing on some more normative aspects of AI-powered personalisation within behavioural science. For now, personalisation will be set aside and other topics discussed. Yet, it is not *gone*, and a reader is encouraged to keep the potential of personalisation in mind throughout the following chapters.

Chapter [3]—Machines Like Us

> "Do you think I am an automaton?—a machine without feelings? and can bear to have my morsel of bread snatched from my lips, and my drop of living water dashed from my cup? Do you think, because I am poor, obscure, plain, and little, I am soulless and heartless? You think wrong!—I have as much soul as you—and full as much heart! And if God had gifted me with some beauty and much wealth, I should have made it as hard for you to leave me, as it is now for me to leave you. I am not talking to you now through the medium of custom, conventionalities, nor even of mortal flesh: it is my spirit that addresses your spirit; just as if both passed through the grave, and we stood at God's feet, equal—as we are!"
>
> —Charlotte Brontë, *Jane Eyre* (2001 [1847], p. 215-216)

The Social Life of Information

Chapter 2 focused on a specific topic—personalisation. This chapter is more thematic. It explores how AI can be used to generate insights which help behavioural scientists respond to difficult problems. These include problems of too much information and obscure information.

That decisions can be impeded by too much information, and that information can be obscured, are not controversial findings within behavioural science. Yet, the manner in which information is discussed within behavioural science often diminishes what one might call the 'social life' of information and encourages one to view information in a way more analogous to computer science—as *bits* of information taking discrete values (e.g., one, or zero). Of course, we all know that information in the ordinary, everyday sense (and therefore, in the behavioural science sense) has a qualitative dimension to it, even if our language often emphasises the quantitative aspect.[82] This chapter will benefit from a brief discussion of these qualitative aspects of everyday information, or, as above, information's 'social life.'

Firstly, many will sympathise with the idea that sometimes one can have 'too much' information. In economics and

[82] One might be tempted to construct some mental model linking information to data, data to knowledge, knowledge to wisdom, and so on. Such models are cute, but do not contribute a great deal to the ideas contained within this chapter. More interesting categories almost certainly emerge when one considers knowledge that can be recorded as information, and knowledge which cannot be (e.g., tacit knowledge; Polanyi, 2005).

psychology, *choice overload* can arise from decisions involving too much choice, and thus too many elements for a decision to consider (Hadar and Sood, 2014).[83] Moon (2010) has somewhat amusingly used choice overload and 'too much' information to explain why supermarkets often sell dozens of different brands of bottled water. Moon argues that by presenting too much choice, consumers become overwhelmed and incapable of making *any* choice, leaving them vulnerable to savvy salespeople who 'help' consumers by 'lending' them their 'expert' knowledge.[84]

'Too much' information also links to Miller's (1956) famous 'magic number seven' paper, which suggests the average person can remember—and effectively manipulate—seven pieces ('bits' or 'chunks') of information in their short-term memory, plus or minus two.[85]

While there is certainly a quantitative aspect to 'too much' information, all information *counts*. For instance, some choices will be readily eliminated from the choice set. It is not that people struggle to navigate many choices *per se*, but that those choices which one already struggles to distinguish may be harder to distinguish when there are many more of them.[86] The inequality of

[83] Though, one review suggests that when and how choice overload manifests is quite complex, and that the broad notion of 'choice overload' is a simplification of matters (Scheibehenne *et al.*, 2010).

[84] Naturally, Moon (2010) implies that this often ends in consumers being upsold. Such effects may be consequential when one is not considering bottled water, but pricey items, such as televisions and cars.

[85] In the medical realm, Obermeyer and Lee (2017) argue that as a result of people living longer, and given ever-increasing medical knowledge, the average patient has become much more medically complicated than in previous decades. More indicators must now be considered, for there are now more known indications of conditions. Meanwhile, patients are more likely to have multiple conditions, and to be being treated with a multitude of medicines. Obermeyer and Lee contend that many medical cases today demand medical professionals utilise much more information than the human mind is capable of working with. Debatably, this is not an instance of there being *too much* information, insofar as this phrase implies that some information is unnecessary. Rather, that in some instances, the amount of necessary information exceeds human cognitive abilities. This is a theme this and subsequent chapters will explore.

[86] I have developed a working hypothesis from many instances of my partner and I being unable to decide what to eat for dinner. When the choice had been narrowed down to two options, I would very deliberately toss a coin to decide between them. I noticed that by choosing randomly, my partner would immediately decide what she *actually* wanted to eat (either through overruling the coin toss, or enthusiastically agreeing with it). My hypothesis was (and is) that prior to the coin toss, there is an uncertainty to the utility (for lack of a better word) of each option, and that these utilities substantially overlap to the point that it is cognitively difficult to distinguish the best option. However, once a 'decision' is made (by the coin), one no longer evaluates the choices in terms of what they *could* have, but in terms of what they must *give up*. This endowment effect switches the choice from a gain to a loss, and we are much more sensitive to losses, collapsing the uncertainty and crystallising our preferences more fully (though still perhaps imperfectly) in our minds. Perhaps one could liken this idea to the quantum uncertainty principle.

information puts an important spin on Miller's magic number seven, also. As Simon (1981) notes, the three-letter string QUV may require three 'chunks' of memory to remember, but CAT is likely to only require one, as is ONE and, ironically, TWO. These latter strings are the same as common words we already know, and thus can draw on associations to remember these words, rather than short-term 'chunks.' QUV, by contrast, lacks those associations, requiring us to remember, separately, 'Q', 'U', and 'V'.[87] This belies another issue—sometimes it is not that information *must* be considered, but that information distracts one from what *ought to be* considered (Simon, 1997a).

Second is the matter of obscurity. Often desirable information is available, but not readily. This can encourage one to seek out *more* information as an unhelpful attempt to remedy the problem (Simon, 1981). Such is often the case in business and government (Drucker, 2006). This problem is tied to that of too much information insofar as too much information can obscure and distract from relevant information, as immediately above. But the lack of easy availability of information is a wider category of impediment which undermines effective decision-making. For instance, the state can feasibly gain information on essentially any topic or domain. But some projects will face substantial budget costs and difficulties from citizens (perhaps because of budget costs) which will disincentivise the undertaking of the project, and the collection of relevant information (Kingdon, 2003). There is also the matter of *ambiguity*, which is taken to mean a scenario which can be explained by two wholly contradictory sets of facts. In ambiguous situations, a lack of availability obscures what information is available by undermining *understanding* of that information (Ellsberg, 2017).[88]

A fun variant of this idea comes when one does not specify what heads and tails mean. Simply toss a coin and ask the other person what the outcome means after-the-fact. To my surprise, many people do not seem to notice that *they are wholly responsible for the decision* in this scenario. We might call this a 'phantom endowment', where the coin toss prompts them to imagine the outcome which they are most sensitive to (either the one they *really* want, or the one they *really do not* want), but which they could not previously appreciate.

Note that once I explained these (untested!) ideas to my partner, she forbade me from using these 'tricks' on her in the future.

[87] Technically, one could recognise that in English, 'Q' is always followed by 'U', and thus 'QU' could be thought of as a single chunk.

[88] Ellsberg (2017) has defined ambiguity this way in his discussion of nuclear war planning. Say a nuclear bomber pilot is ordered to drill a nuclear attack plan. Wanting an accurate test, commanders do not tell pilots involved that it is a drill. The diligent pilot loads up their nuclear payload, taxis along the runway, and proceeds towards the enemy city. The pilot knows that if this were a test, they would soon receive disarm orders from the base. Then, suddenly, there is a flash of light, followed by an enormous explosion. A nuclear blast on the base that the pilot has just taken off from. Ellsberg invites us to consider what the pilot should do. On the one hand, the pilot has been told nuclear war has begun, has

While one might quantitively understand information within an organisation or within the brain of a decision-maker as a 'bit' represented by the movement of an electron around a circuit board, one might also understand it in more qualitative, analogue terms, say as a body of water or mass of sand filling up and draining from a leaky system of buckets and pipes. Herein one stumbles across a tension at the heart of AI as an information management tool within behavioural science and beyond. Some applications of AI may compliment the 'social life' of information, say by empowering experts to make better decisions. Others may act as if the 'social life' does not exist, leading to challenges, often as the social life is forced to adapt to the mechanical philosophy of the AI system. This chapter will present examples of both.

Filtering Information and Finding Biases

Perhaps the most important application of AI within behavioural science is as a data analysis tool.[89] AI, and specifically predictive AI, presents novel ways of analysing data, and encourages the uses of different (and more) data. In doing so, predictive AI may support behavioural scientists both in their investigations of human behaviour and in their deployment of these insights to problems in everyday life.

Aonghusa and Michie (2020) discuss a fascinating application of AI in their work as part of the *Human Behaviour Change Project*. An enormous amount of literature exists examining various interventions to change habits and encourage healthier lifestyles. The amount of literature alone might be daunting, but the variety adds an additional dimension which creates informational challenges.[90] Different studies use different samples and sample sizes. They target different behaviours via different behavioural mechanisms by deploying different behavioural interventions. Some are experimental studies, others quasi-experimental, and so on. The heterogeneity of the literature frustrates one's abilities to draw coherent insights from the

been ordered to attack the enemy, and has just seen a nuclear explosion. *It all makes sense*. On the other hand, the explosion might have been caused by an accident involved as part of the drill. The blast has now killed the commanders who would have issued the disarm orders. War has not been declared, and the worst thing the pilot can do is to keep going. *It all makes sense*. Ellsberg points to the inherent ambiguity of this situation: two sensible but contradictory interpretations of the same information. The pilot's ignorance of what caused the blast obscures how they should interpret the information that is readily available to them.

[89] It is not unreasonable to understand AI personalisation simply as a kind of applied data analysis.

[90] Aonghusa and Michie do not give details as to the size of the corpus on which their model is trained. My estimate would be several thousand papers, given the ever-increasing amount of research being undertaken around the world, and in behavioural science.

literature. It also undermines adaptiveness to real-world policy pressures, from unfolding public health crises to political changes which reorientate priors, budgets, and philosophies on public service.

The programme of work discussed by Aonghusa and Michie involves using an AI system to synthesise these many thousands of public health studies. Then, through a user interface, researchers can query the system for predictions based on the literature. As queries change, in response to the changing policy environment, the system can make new predictions, which policymakers can use to update their plans. Arriving at a synthesised prediction of a public health intervention and updating the prediction as circumstances around the intervention change, would be tremendously difficult for even a trained team of researchers to accomplish. Aonghusa and Michie note that the use of AI for these functions empowers public health experts to better utilise valuable research *which is available*, but which is not *readily accessible*. In a comparable and recent study, Kaiser *et al.* (2024) report that a fine-tuned large language model (LLM) can demonstrate a high predictive accuracy of behavioural experiments to change eating habits. This leads these researchers to suggest that, through further refinement and technological development, AI could become a vital information management and prediction tool for behavioural scientists and policymakers. Luo *et al.* (2024) report similar results from LLMs trained to predict neuroscience results.

In these instances, AI does not replace human judgement, or even behavioural science expertise. Neither Aonghusa and Michie nor Kaiser *et al.* and Luo *et al.* suggest that such AI applications can function successfully without a skilled group of behavioural practitioners. The predictions discussed by Aonghusa and Michie simply attenuate practitioner knowledge and support a final, human-determined policy decision. Kaiser *et al.* and Luo *et al.* emphasise how LLMs must be fine-tuned by behavioural practitioners to reliably predict study outcomes. Naïve models, as well as an *overly* tuned models, may fail to make predictions with an adequate degree of accuracy. Furthermore, in these studies, people are required to undertake the behavioural research from which the system is trained—in the case of Aonghusa and Michie—or compared—in the case of Kaiser *et al.*, and Luo *et al.* These studies thus approach AI systems as *tools* for researchers and practitioners.

Another area of AI application is likely to be in detecting behavioural biases (Mills, Costa and Sunstein, 2023). While this remains an area dominated more by speculation than practical results, there are robust arguments and *some* results which suggest this is a feasible application to consider.[91] Mills, Costa and Sunstein (2023) argue that one may understand a behavioural bias as a tendency within data. Consider the default bias. One might imagine that a dataset showing two choices, a default option (coded as '1') and an alternative option (coded as '0'), approximately half of observations would be registered as a '1' and half as a '0'. A significant deviation from this pattern, favouring '1' over '0', would be indicative of a bias towards the default, assuming the default option changes arbitrarily for different people, and assuming several instances of this experiment (with different choices, participants, and environments) demonstrate the same result. From this perspective, behavioural biases will arise as *patterns* in datasets. Predictive AI technologies, by design, spot and act upon patterns. Thus, the argument is that once behavioural biases are understood as patterns in data, AI technologies may help behavioural scientists identify known biases in novel settings, and perhaps more interestingly, new biases in old and new settings alike.

Some work around natural language processing (NLP) attests to this hypothesis. *Word2Vec* is a relatively old text analysis approach drawing on some AI techniques to gain insights from big text data which would be extremely difficult for people to manually calculate (Mikolov *et al.*, 2013). The technique *vectorises* words, essentially allowing words to be represented as vectors of numbers. These vectors numerically encode semantic relationships between words. This could, *in principle*, be undertaken by a team of people, but with relative delay and difficulty given the mass of text which one might wish to analyse. As words come to be represented as numbers, the relationships between words can be analysed mathematically. A common analysis is to examine the semantic similarity between different words.[92]

[91] Note, these arguments come from myself and my colleagues. I may be biased as to their strength and credulity.

[92] *Word2Vec* was the most prominent of a group of word vectorisation applications developed in the mid-2010s. Ideas found within these models contributed in part to the development of today's large language models, though ideas such as 'attention' within the transformer architecture. Developed by Google engineers, *Word2Vec* essentially predicts patterns between words using a relatively basic neural network consisting of a single layer of neurons. Through training, this layer comes to 'embed' numerical information which captures the relationship between different words (without the numbers themselves necessarily meaning anything). This allows individual words to be represented as vectors of numbers (hence *vectorisation*), which can be probed mathematically to infer semantic relationships. For instance, semantic similarity (how similar are two words in everyday language) can be measured by observing the 'distance' between the vectors of the respective words in an *n*-dimension vector space, where *n* is the length of the vector (this information can then be used to tell an LLM what to 'pay attention to' during the training process—a rudimentary description of how a transformer works).

Consider the words ‘apple’, ‘orange’, and ‘fruit’. Intuitively, both apples and oranges are *fruits*; both words are likely to arise in discussions of fruit, and in some instances may be substitutes for the word ‘fruit’, while neither has any sensible reason to be more similar, or more substitutable, for the word ‘fruit’ as the other. Thus, ‘apple’ and ‘orange’ are likely to have a high similarity to the word ‘fruit’ compared to, say, the word ‘elephant’ (and their similarities to the word ‘fruit’ are likely to be, well *similar*). Yet, the similarity between ‘apple’ and ‘orange’ themselves is likely to be lesser than either to the word ‘fruit’. After all, the notion of ‘apples or oranges’ is often used to invoke two things that are quite different. Still, both being fruits, they are still likely to be quite similar, again comparing either to the word ‘elephant.’ But what about the word ‘iPhone’? In this instance, ‘apple’ is likely to be much more similar to ‘iPhone’ than ‘orange’ will be. Likewise, the word ‘paint’ will probably be more similar to ‘orange’ than to ‘apple.’ And so on. Similarity, in all these instances, means co-occurrence and substitutability of terms in natural language, with relative distances encoding different semantic contexts (e.g., the closeness of ‘apple’ and ‘orange’ to ‘fruit’, the closeness of ‘apple’ to ‘iPhone’ and ‘orange’ to ‘paint,’ and the distance of both from ‘elephant’).

Methods for realising meaningful results from text data are already available (e.g., various qualitative methods), and one should not regard the development of word vectorisation (or word *embedding*) techniques as a novelty insofar as it allows behavioural scientists to gain insights from text data. Where these techniques are novel is in how they enable behavioural scientists to analyse massive amounts of text data; volumes of data which may exceed the practical use of some qualitative methods. In this sense, it is important to appreciate this AI application as still being one which falls into the information management and data analysis category. How can behavioural scientists use these techniques to identify biases?

One immediate insight can be gained from looking at the outputs of these models once trained on relevant data. Systems like *Word2Vec* has been found to often encode gender-biased word associations (Bolukbasi *et al.*, 2016; Brunet *et al.*, 2019). Such results suggest that, in everyday language, people may explicitly or implicitly use a particular corpus of words when describing men, and a different—and, often, less favourable—corpus of words when describing women. There are limits to how much one should extend the notion of word associations capturing real-world biases. If one rejects the rationale behind word vectorisation (e.g., that probabilistic co-occurrence is indicative of some semantic relationship), then such gender-biases may be more indicative of choices around model design, rather than of biases attributable to the people whose language makes up the underlying dataset. Equally, if one accepts the rationale of the model, a sexist word association within the model may be indicative of gender biased behaviour within the data, and thus within the population from which the data is taken.

This latter argument has been extended into the development of 'word embedding association tests' (WEATs), where word vectorisation is used to investigate implicit biases in groups, before being compared with implicit biases identified through established behavioural science methods, most notably the implicit association test (IAT). Caliskan *et al.* (2017) develop the WEAT approach and demonstrate that it can accurately replicate the implicit biases found in a separate IAT. In another study, Charlesworth *et al.* (2021) use word embeddings to demonstrate how gender stereotypes persist in big text data, showing how associations between gender and toys, media, occupations, and so on, follow expected patterns of stereotyping. Evenepoel (2022) extends the WEAT method further. By analysing big text datasets from different decades, Evenepoel (2022) shows how biases and attitudes have changed over time. For instance, one WEAT of attitudes towards depression demonstrates a declining association between depression and psychological causes, perhaps reflecting greater appreciation of the various causes of mental health, compared to popularly held beliefs in the mid-twentieth century.[93] It has even been shown that word embeddings can accurately predict the ideological positions of politicians by identifying common word associations found within different parts of the political spectrum (Rheault and Cochrane, 2020).

Word embeddings extend the behavioural science toolkit in several ways (Feuerriegel *et al.*, 2025). The approach offers an alternative to methods such as the IAT, which might appeal to the preferences and expertise of some practitioners. The big text data angle also allows one to go beyond the scope of IATs, say by introducing analyses of biases *over time*, something which would be difficult to achieve through an IAT.

Alternate to word embedding approaches, others have used AI methods to identify a variety of data points which might subsequently be called biases. Various studies have investigated how doctors and judges make decisions around, say, referring potential heart attack patients for further testing, or determining whether a defendant should be offered the opportunity for bail, respectively. By training predictive AI models on various data

[93] Evenepoel (2022) notes that data from more recent decades does not show associations which are statistically significantly different from the 1950s, but the trend away from a depression-psychological association is evident.

available to these decision-makers, researchers can then probe these models to determine which data points have an outsized predictive power within the model. Assuming these especially powerful variables do not conform to what one might consider 'rational' standards for decision-making, the predictive power of the variable may be indicative of biased behaviour on the part of the decision-maker (Ludwig and Mullainathan, 2022, 2021).[94]

For instance, evidence suggests that judges frequently exhibit two biases when making bail decisions, which Sunstein (2023) dubs 'current offence bias' and 'mugshot bias.' As these names suggest, studies find that a defendant's current offence has an outsized role in whether they receive bail, when—one might argue—their *whole* offending history is a better indication of whether one is likely to commit another crime if granted bail (Kleinberg *et al.*, 2018). In the case of the mugshot, a defendant's mugshot has outsized predictive power when predicting if bail will be granted (Ludwig and Mullainathan, 2022), when a person's appearance should not have any bearing on their likelihood to commit a crime in the future.[95] Sunstein (2023) argues that both have links to the more 'fundamental' bias of availability bias (Tversky and Kahneman, 1974)—the tendency to focus on immediately available information (like a mugshot, or current offence) rather than less apparent information (like a defendant's past offending history).

A similar such conclusion can be reached when considering Mullainathan and Obermeyer's (2022) AI-led investigation of medical referral procedures. They find that doctors tend to over-test low risk patients (sending them for a battery of tests which

[94] Rational is used here for lack of a better word. One might say 'reasonable' or 'intuitive' instead. One might also defer to, say, professional guidelines which offer best-practice advice for decision-making. If models suggest data points *not* found in these guidelines are used, one may be justified in calling this a bias (Ludwig and Mullainathan, 2021).

[95] These findings relating to justice can be contentious. For instance, one may suggest that the current offence bias should not be considered as such. Assuming the criminal justice system serves its function of rehabilitating, and that once a punishment has been served a citizen is no longer bound to atone for their crime, then one could ask if it is not unfair to consider a defendant's whole past history? What is called a bias, in this instance, is merely a matter of perspective about the function of the criminal justice system. Chapter 4 will elaborate further.

In both instances, it is also worthwhile to ask whether labelling these 'biases' as such is actually helpful. Kleinberg *et al.* (2018) argue that biased judicial decisions lead to inefficiencies in sentencing, with low-likelihood reoffenders being jailed and high-likelihood reoffenders being granted bail. Thus, they contend, tackling these biases can have substantial social welfare effects. Nevertheless, such advocacy neatly sidesteps the more pressing questions of *what drives offending?* More efficient allocation of prison resources is hardly efficient if equal or greater effort is not being shown to reducing the *demand* for prison, so to speak. A more-efficient allocation of a wasted resource is still waste. To this end, it is perhaps helpful to reconsider some of the discussion in Chapter 2 around personalisation. There, it was noted that personalisation could end up reproducing unjust, discriminatory, practices.

would determine severe illnesses they are unlikely to have) while under-testing high risk patients (who should be subjected to many tests which, compared to the average, might seem excessive). This, Mullainathan and Obermeyer report, leads to inefficiencies in healthcare provision, with many patients subject to waste (in terms of time and medical costs), and some subject to risk from undiagnosed or underdiagnosed symptoms. They also note that rational precautions—say, because doctors do not want to be sued—do not explain the pattern of under- and over-testing. Instead, doctors seem to evaluate a patient's health risk based on immediately available information, rather than a more complete medical history and circumstantial details which would inform a more accurate assessment of risk. As with judges, this finding is the result of AI-based analysis techniques which allow influential decision-making factors to be investigated.[96]

Such insights can inform policy recommendations. For instance, judges could receive algorithmic recommendations as to whether a defendant is likely to reoffend, while doctors could be told of a patient's risk, as predicted by an AI model. The formatting of the information judges receive might also be reconfigured so that the most relevant information (in terms of best guidance) is also the most available information, while irrelevant information (like the mugshot) is less available, or even removed entirely (Ludwig and Mullainathan, 2021; Sunstein, 2023). So too might medical information. One worthwhile criticism of such recommendations, though, is the potential for AI recommendations to gain an outsized influence in decision-making processes which should be undertaken by a person.[97] To this end, one might return again to Simon's (1987a, 1987b) advocacy for

[96] One recent study which combines both natural language analysis through AI methods with attempts to identify biases in expert judgements, like doctors and judges, is given by Jelveh *et al.* (2024). They trained an AI algorithm on the natural language of economics papers to predict the political leanings of academic economists. Then, using these political predictions, they examined the spread of policy recommendations. Assuming economists analyse data objectively and using the same robust methods, one would anticipate relatively small variation in policy recommendations. However, this study suggests that left-leaning economists advocate, on average, for a top tax rate which is 14% higher than right-leaning economists.

[97] Later chapters will expand more on this. Human oversight is not infallible, and Sunstein (2024) has suggested there may be instances where algorithms should make decisions, though, with the caveat that humans could choose to overrule those decisions. This is politically contentious—for instance, one could imagine that when the criminal justice system is under pressure, lacking enough judges and economic resources, the 'choice' to defer to an algorithm becomes only nominal, and political reality implements a *de facto* justice-by-algorithm court system. Likewise, where doctors face legal suits for poor medical treatment, it may often become safer to *always* defer to the algorithm, rather than overrule it, knowing that if one is wrong, the case against them will seem compelling.

human-AI collaboration and AI as an 'expert support system.'[98] Here, insofar as AI can identify and remove distracting information which undermines the effective decision-making of a judge or a doctor, without the AI *actually* making the decision, may allow a happy medium to be reached where a human decides, but decides more consistently and equitably.[99]

Who Are You Talking To?

Online choice architecture (OCA) is a behavioural scientist's way of talking about a website, or what a user interface (UI) designer would call a *user interface* (obviously). This is not to suggest that OCA represents a proliferation of terms for its own sake, as if the behavioural scientists cannot play nicely with the UI designers. The utility of a term like OCA is it encourages one to see interfaces that are interacted with every day as drivers of individual choice rather

[98] It is interesting to note that, in the medical field (and perhaps others), Simon's perspective was hardly original.

As early as 1960, Lusted (1960) argued that computers could function as information filtering systems in medicine, splitting suspect patient scans off from non-suspect ones: "an electronic scanner-computer [could] look at chest photofluorograms, to separate the clearly normal chest films from the abnormal chest films. The abnormal chest films would be marked for later study by the radiologists." According to Greene and Lea (2019, p. 480), Vladimir Zworykin (who also invented the TV) had "warned that medical data were accumulating at a pace exceeding physicians' cognitive capacity" as early as 1964. This followed perhaps one of the first information revolutions anywhere, with new technologies for diagnosis and information storage leading to rapid increases in medical data collection throughout the 1950s.

Interestingly, much of the concern about 'judgement' was not directed at *doctors'* judgement, but at the judgement of *programmers* of computers. This is not wholly absent from discussions today (e.g., Mullainathan and Obermeyer, 2017). Yet, recent behavioural science studies have emphasised the fallibility of *doctors* and the benefits of AI algorithms as a result. This is perhaps a meaningful inversion of what is—to reiterate—*hardly a new area* of discussion. I have argued that this inversion may be understood by examining the relatively 'narrow' view of bounded rationality popular in behavioural science today, compared to Simon's (1997a, 1981) initial discussions of the idea (Mills, 2024).

To be sure, there was recognition of doctors' bias in the sense that doctors might disagree, preferring information which aligned to their specialty (Ledley and Lusted, 1959), though this is hardly a deviation from Simon's (1997a) observations in the 1940s about information preferences in organisational decision-making. This initial emphasis on biases and disagreement remained orientated towards the challenge of *programming* a useful AI for doctors. Such observations were not necessarily used as *motivating* the use of AI technologies, as one might argue is the case today. As Greene and Lea (2019) note, Zworykin was concerned about such a conclusion and felt it necessary to emphasise that the goal was not to undermine doctors. Neither, do I think, is the objective of recent behavioural science studies. An emphasis on *doctors'* biases, nevertheless, risks doing so.

[99] This 'happy medium' approach to algorithmic involvement would probably find support amongst individuals like Sunstein. Firstly, because Sunstein (2023) accepts that some expert decision-makers do indeed outperform predictive AI algorithms. Secondly, because the suggestion that experts should be given AI recommendations could *also* be interpreted as removing distracting information and improving the quality of information available to decision-makers.

than merely neutral receptacles of options. Much like 'analogue' behavioural science did not necessarily propose radically different ideas to those found in marketing, but did emphasise a different viewpoint from which insights could be gleamed, so too with 'digital' behavioural science and OCA.

The OCA space is growing quite considerably. To an extent, the intersection of AI with OCA is driven more by the pre-existing interest in technology found amongst OCA researchers, than by compelling evidence of the role of AI. Though, this is not a universal statement. Recommendation algorithms are a common component of OCA, typically found in a 'recommended for you' box on retail websites.[100] Unpicking designs and highlighting the behavioural science contained within them can also be time consuming and subject to human bias. Automating such 'auditing' processes may also be a domain where AI could supercharge OCA research and support regulatory goals of protecting customers and fostering greater competition amongst firms (Mills and Whittle, 2024). But perhaps the most interesting intersection involves chatbots powered by generative AI. This is a nascent area—one which may slip more into speculation—but of importance given the growing experimentation with generative AI and given existing OCA which perhaps makes AI sales assistants useful.

It is helpful to understand the emerging role of AI chatbots in OCA as an evolution of recommendation algorithms, or even search bars. Both recommendation algorithms and search bars respond to the problem of too much information, and obscure information, by providing users of websites with more behaviourally accessible means of navigating what the website has to offer. The search bar allows a user to enter a loose term which they think will be relevant; the recommendation engine predicts what the user might already consider relevant.

The AI chatbot can be understood as serving a similar purpose, at least in principle, while extending the capabilities through which a person can interact with a website to include natural language communication. Through natural language interactions with users of a website, an AI chatbot could serve as a valuable information management tool. A person could ask the chatbot to recommend products on a retail website, or to provide a link to an online form for claiming some government provision. The UK Government (2023), for instance, has outlined in its guidance to civil servants that the use of large language models (LLMs) as part of supporting citizen queries could be an appropriate application of AI. Its Government Data Service (GDS)

[100] Readers are referred to Chapter 2, where the discussion of personalisation has dealt quite substantially with the role of AI and behavioural science in recommendation algorithms, even if somewhat indirectly.

has argued that "there is potential for [AI chatbots] to have a major, and positive, impact on how people use [government websites] – for instance making it easier to find answers to their questions from the 700,000+ page estate" (Bellamy, 2024, para. 3). A subsequent experiment run by the GDS found around two-thirds of users to be satisfied with an early AI chatbot on government webpages (Gregory *et al.*, 2024).[101]

The relative recency of consistent, high-quality generative AI chatbots is reflected in the relative scarcity of robust tests demonstrating the behavioural impact of AI chatbots as part of OCA. One study in this space comes from the Behavioural Insights Team (2023). In partnership with the UK Government, they investigate various behavioural outcomes from citizens using AI chatbots to navigate government websites. The findings point to a rather more complicated picture than simply that AI chatbots help people navigate information better.

Firstly, those given access to an AI chatbot performed worse on a multiple-choice task than a control group with no AI chatbot.[102] At the same time, only around 40% of those with access to the chatbot actually used it. One explanation of the poor performance may therefore be that for most people, the chatbot window was merely an additional distraction adding to the informational complexity to be navigated. This is perhaps demonstrated through a second finding.

Secondly, of those who *did* use the chatbot, accuracy did in fact improve in some instances.[103] Interestingly, those who used the chatbot were also *slower* in completing the multiple-choice task—though by only around three seconds. Experientially, this and the use of an AI chatbot in general appears to have had a relatively minimal impact. Participants did not really find government information any more or less difficult to understand, though perceptions of task ease were higher in three of the four treatment groups, compared to the control.

[101] While acknowledged as only a preliminary study as part of a wider scheme of work, the relatively low sample size (N = 157) and vague language (e.g., "satisfied") means it is difficult to extrapolate too much from this result.

[102] Participants were asked questions about government policies relating to housing and health. Such information could be found on government websites. This result suggests that the chatbot *negatively* impacted a person's ability to locate relevant information.

[103] The Behavioural Insights Team (2023) tested four different AI chatbot designs. The most intrusive designs (whole page chatbots) lead to worse performance or identical performance compared to the control group who had no access to an AI chatbot. The less intrusive designs saw improvements over the control, though it is not clear if these improvements were statistically significant.

Others have found similar results which point to a mixed bag and the need for additional research in this space. Aoki (2020), for instance, finds that public trust in AI chatbots on government websites is significantly dependent on *what* information the chatbot is being used to find. Chatbots tend to be trusted more on recycling guidance than on parental support, for instance. In the retail space, Blut *et al.* (2024) echo the Behavioural Insights Team insofar as they find that the most persistent challenge around AI chatbots is not necessarily accuracy or trust *per se*, but simply *getting people to use them*. Perhaps such a challenge will erode if AI services become more integrated into everyday life, such as in online shopping and on government websites. But until such erosion occurs, there is a risk that rather than tackling too much information, revealing the obscure, and cutting through distractions, AI chatbots *contribute* to these problems.

While retailers may share some of the challenges and opportunities of AI chatbots with governments, they also face considerations unique to the private sector. One recent study (Castelo *et al.*, 2023) has found that consumers dislike AI chatbots compared to human customer service attendants, in part because consumers believe that private firms only use AI for their own private benefit, such as being able to cut down on staffing costs. The same study has found, though, that consumer attitudes improve when the AI chatbot produces a clear benefit for consumers, too, either through lower prices (e.g., discounts that only the chatbot can offer) or superior information (e.g., a significantly improved customer experience). This being so, there is hope that adoption of AI chatbots may improve—though it is questionable to what extent AI, rather than other improvements in online services, will encourage this adoption.[104]

Such findings belie an important aspect of AI chatbots which should return attention to older ideas about sales and persuasion. An AI chatbot placed on a retailer's website is analogous to a salesperson in a store. While both *may* be helpful (they are both typically called 'assistants,' after all), both the AI chatbot and the salesperson work for the retailer and are employed precisely because it is believed they increase sales.[105] Though, it is not always

[104] Something to note here is the distributional effects of AI chatbots. It seems reasonable to suspect that most customer service complaints are relatively similar and concern a small set of problem areas. For most people, a generative AI chatbot can probably support the efficient resolution of their issues. But for a minority, issues will be complicated, unusual, and perhaps wholly *novel.* These issues may test the efficacy of an AI chatbot, and should the chatbot struggle, may *exacerbate* rather than *assist*these particular customers.

[105] At the time of writing, there does not appear to be clear empirical evidence that AI chatbots have a causal relationship with higher sales. Though, there is substantial discussion of AI chatbots within some industrial magazines and

clear *how* this is achieved. Maybe both improve the customer experience, which indirectly increases sales by, say, encouraging customers to return more frequently than they would if they *just* liked the product? Or maybe both effectively deploy persuasion techniques to encourage the customer to purchase a product that they otherwise might not?

Such exercises in speculation are what marketers and consumer behaviour researchers call *metacognition*—thinking about thinking (Friestad and Wright, 1994; Wright, 2002). Metacognition is an activity we all engage in, and a frequent example is when one walks into a store and must have a conversation with a salesperson. In such conversations, neither 'agent' (the customer or the salesperson) knows the exact motive of the other. These goals are likely different but may align in some aspects. For instance, a customer who wants a high-quality product may be satisfied by a salesperson who sells them an expensive product, assuming a loose correlation between quality and price. Metacognition enters the fray insofar as each agent is trying to determine what the other agent's objective is, so they can adopt appropriate strategies to either persuade an agent (from the salesperson's perspective) or to defend from persuasion (from the customer's perspective).[106]

Now replace 'salesperson' with an AI chatbot. The first benefit of doing so is one may gain a new perspective on the question of trust of AI. The degree to which people trust AI is likely influenced by the context in which the AI chatbot is deployed. For instance, an (ideal) government acts in the interests of its citizens. As such, metacognitively a user might determine that regardless of the chatbot's capabilities, its objective *is to help them*, leading to higher trust. In a different context—online retail, for instance—it is reasonable to assume adversarial objectives between the chatbot and the user.[107] Indeed, even in the governmental context, this might explain why Aoki (2020) finds trust in AI chatbots to be influenced by the policy context in which the chatbot is deployed.

Where this metacognitive perspective becomes interesting is around the question of knowledge. Friestad and Wright (1994) argue in their model of metacognition that knowledge is a key driver of metacognition, and thus successful persuasion.

amongst various consultancies. For the purposes of discussion, I will assume that firms at the least *believe* AI chatbots positively impact sales.

[106] Technically, this is a description of *marketplace* metacognition. The notion of 'thinking about thinking' is quite broad, and is often applied to non-adversarial scenarios, such as when a student evaluates their own learning, or a teacher tries to understand why a student is struggling to learn.

[107] By 'adversarial' I do not mean in total conflict. Instead, adversarial should be understood as objectives where it is possible for both agents to succeed, or for one agent to succeed while the other fails.

Knowledge takes different forms. For instance, someone who is very knowledgeable about a product will likely discern the persuasion strategies of a less-knowledgeable salesperson, as the salesperson may advocate for a product the knowledgeable consumer knows to be inferior (Moon, 2010). Another form of knowledge is knowledge of persuasion techniques themselves. Salespeople will often encourage wavering customers to think about how they would use a product once they have purchased it ('call to action'). They might recount the 'testimony' of other satisfied customers ('appeal to conformity'). A customer's knowledge or ignorance of such techniques is likely to influence their metacognition insofar as they can identify these techniques or not.[108] Finally, and perhaps most interestingly, how much agents knew about one another—so-called agential knowledge—influences metacognition. For instance, simply knowing that one agent is a salesperson may prompt the other to be more wary of persuasion.

AI chatbots are likely to have superior product knowledge than the average user. This is partly why they are deployed in the first instance, and so should not considered a major cause of concern. More relevant is how an AI chatbot could create knowledge asymmetries between itself and an interacting customer, in terms of persuasion knowledge and agential knowledge.

A model trained on a large body of text will likely be able to demonstrate superior persuasion knowledge for similar reasons to its superior product knowledge. But such a model may also be able to adapt its use of such techniques in response to ongoing dialogues with users.[109] For instance, *mirroring* is a persuasion technique which involves matching the behaviour and personality style of the opposing agent. Marketing research suggests people respond more positively to sales strategies which match their personality, making mirroring an effective sales technique (Moon, 2002). Murphy (2024) finds that LLMs like GPT-3.5 and GPT-4 can predict individual personalities from text data with around a 75% accuracy.[110] They note that such capabilities could be used to

[108] Incidentally, this is part of the reason why some items require a deposit to be paid or allow one to pay in small parts. While metacognitive selling is not the whole story (deposits can pay for work to be done before final delivery; part-payment can generate greater income than a one-time payment), these techniques form part of a whole package of techniques designed to convey advantages onto the retailer.

[109] This would be a form of personalisation. See Chapter 2.

[110] Murphy (2024) uses the 50 most recent tweets of study participants. Precisely how much data is needed to determine personality is an important question, given low user engagement with AI chatbots (at present), and given the likely short conversations those who use chatbots actually have with them. For instance, if very little text can accurately predict personality, then a retailer might prioritise *any* engagement over pro-longed engagement—especially if tied to a customer account, allowing a long-term profile to be assembled. If longer

influence people. One immediate application might be for an AI chatbot to mirror the personality of the customer it is engaged with.

While, in principle, a skilled salesperson could replicate the superior persuasion knowledge and adaptability of an AI chatbot, one might speculate that AI chatbots can more reliably establish this information asymmetry, and thus advantage in terms of metacognition. A similar situation might be hypothesised around agential knowledge.

When communicating with an AI chatbot, users have a powerful piece of agential knowledge—they know the chatbot is, well, a chatbot. But they will likely know little else which might provide them with a strategy for metacognitive defence.[111] By contrast, AI chatbots could be designed to incorporate a huge array of agential knowledge about the user. Some knowledge might be derived—as in the case of personality—but other knowledge may already be held by the retailer and put to use by the chatbot.[112] For instance, the purchase history and location of a previous customer will likely be known to a retailer. Such information might then be used to develop persuasion techniques against which a customer is more vulnerable. This is not necessarily a *negative* application—the use of superior agential knowledge of the customer might allow the customer to achieve superior outcomes—but in an adversarial, metacognitive context, such an application is likely to be deployed primarily for the maximal benefit of the retailer. For instance, if a retailer knows (approximately) when a spouse's birthday is, an AI chatbot might *naturally* incorporate a reminder to buy them a present into an unrelated sales chat.

conversations are needed, a retailer might be incentivised to prolong conversations, leading to different prompts and designs of chatbots. If mirroring personality does not significantly increase sales, such adjustments may not be implemented. Though, this is not necessarily the only reason why a retailer might seek to retain customers in chatbot conversations.

[111] It is difficult to disentangle, at present, to what extent people avoid using chatbots because such devices have yet to be widely adopted, or because people are acting in a way one might describe as metacognitive. In both instances, a rationale for why one might wish to pass an AI chatbot off as a 'real' person could be made. From a metacognitive perspective, doing so would just compound the asymmetrical advantage of agential knowledge. Rather than, perhaps, thinking it is creepy how much an AI knows about you, one might instead conclude that the 'human' online sales assistant is really attentive, and really 'gets you.'

[112] One could speculate about other 'latent' behavioural datapoints which could be collected through interacting with an AI chatbot. For instance, the time it takes a user to reply might be indicative of their engagement with the conversation, their level of (im)patience, and so on. Analysis of formatting, grammar, and complexity of language might be used to estimate education level, or, again, the hurriedness of the customer. All potentially represent agential knowledge through which persuasion strategies could be developed (e.g., a hurried customer is likely less sensitive to price than a customer with ample time to browse).

Economically, it is interesting to note that AI chatbots enhance the value proposition of sales. A salesperson will likely discard much of the agential knowledge they have gained about a customer once the customer leaves the store—particularly if the customer does not buy anything. An interaction with an AI chatbot, by contrast, leaves a data trail which firms can mine for greater customer insights and, ultimately, sales (Zuboff, 2015). While mere speculation, one might hypothesise that the optimal deployment of an AI chatbot (in *sales*) is not whatever allows the customer to find what they are looking for promptly, nor even whatever encourages a customer to buy a more expensive product, but what allows a retailer to extract as much agential knowledge from the customer as possible while still maximising sales.[113] From an OCA perspective, which is often concerned with consumer welfare, this and above speculations raise important questions which should elevate the study of AI chatbots to the forefront of researcher's minds.

Though, the use of AI chatbots to gather data and discern insights need not always be deployed for the exploitation of consumers. A fascinating study by Chopra and Haaland (2023) demonstrates how AI chatbots can be used to conduct qualitative research at scale. Qualitative research, such as interviews, typically deal with small scales in comparison to quantitative studies. This is because it can take a great deal of time to find participants, arrange, and then conduct interviews. Qualitative scholars who dedicate years to a subject and bring their results to the public in, say, a book, might have only conducted several hundred interviews in that time. Chopra and Haaland (2023) argue that the use of AI chatbots in qualitative research may allow such methods to reach scales comparable to quantitative research. In their approach, established research participant recruit tools are used to reach several hundred interviewees, who then engage in an interview with an AI chatbot prompted on key questions related to the research.

Chopra and Haaland (2023) champion this approach because it may allow qualitative researchers to engage in topics and fields which have typically shunned qualitative methods, like economics. Such an ambition is admirable. Nevertheless, the application of AI chatbots here is hardly perfect. For instance, one reason for small samples in qualitative research is the scarcity of suitable interviewees. An AI chatbot does not overcome this problem.

Digital Clones and Simulated Societies

To this end, it is helpful to consider another application of AI, and generative AI specifically, within behavioural science—*silicon*

[113] Such a hypothesis is unlikely to hold for AI customer support chatbots, as value here is likely maximised by prompt resolution of problems.

sampling. Simulations have been used within social science for several decades, with varying degrees of sophistication (Bonabeau, 2002). Silicon sampling methods seek to exploit properties of LLMs to simulate populations, which can then be sampled and experimented upon in numerous ways; often in ways which for one reason or another would be infeasible on a 'real' sample of people. In this sense, silicon sampling is not a means of tackling 'too much information' as it is a means of *expanding* information access. Many advocates of silicon sampling would align the approach with one of revealing information which is not immediately accessible to people, but is 'embedded' within trained AI models. Doing so, one might argue, can enhance practitioner understanding of those they seek to influence, or design policy for, leading to improved outcomes for all.

At the heart of this idea is the notion of *algorithmic fidelity*. Argyle *et al.* (2023, p. 339), who propose the term, define algorithmic fidelity as, "the degree to which the complex patterns of relationships between ideas, attitudes, and sociocultural contexts within a [large language] model accurately mirror those within a range of human subpopulations." An AI model is high in fidelity if it produces outputs which correspond to those which would be produced by distinct groups and communities. Likewise, low fidelity models may arise through poor prediction of those groups, or through an emphasis on the *average* of the whole population. Argyle (2023, p. original emphasis) contend that emerging AI models demonstrate high fidelity because, "these language models do not contain just one bias, but *many*" allowing AI models to be, "biased both toward *and* against specific groups and perspectives in ways that strongly correspond with human response patterns along fine-grained demographic axes."

Unlike AI chatbots in OCA, where one must be more speculative as to the behavioural effects of these technologies, a relative explosion has occurred in the behavioural science literature concerning silicon sampling since around 2023. One area in which silicon sampling has been readily embraced is in the consumer behaviour literature. Brand *et al.* (2023) use GPT-3.5 and a representative silicon sample of survey respondents to simulate consumer preferences for various products. They report that the simulation produces statistically comparable responses to a human sample of survey responses in terms of consumer preferences and willingness-to-pay. Similarly, Hämäläinen *et al.* (2023) find that GPT-3 is able to accurately simulate accounts of consumer 'experiences' and 'opinions' about products, as measured by the ability to distinguish AI-generated responses from human responses. Hämäläinen *et al.* (2023) emphasise that silicon sampling may be especially useful in consumer behaviour research because of the field's connection with market research, where new products must often be rapidly piloted on an appropriate target market. Such activities can be costly and take longer than competitive deadlines allow. These issues might be resolved through accurate silicon sampling techniques. From an economic decision-making perspective, which also has relevance to consumer behaviour and marketing domains, silicon samples have been found to demonstrate responses to economic games which are comparable to those given by human participants (Aher *et al.*, 2023; Mei *et al.*, 2024).

However, these studies in consumer behaviour do not reflect the full literature on silicon sampling. Studies on political decision-making in particular demonstrate various challenges related to minority representation in LLMs. Lee *et al.* (2023) find GPT-4 can generate synthetic populations which accurately simulate presidential voting behaviours and policy positions.[114] Though, such high accuracy is only demonstrated when GPT-4 is given various contextual priming. Even with such additional prompting, Lee *et al.* (2023) report discrepancies, with GPT-4 under-estimating support for some policies amongst minority groups, such as Black Americans. In a similar study, Hwang *et al.* (2023) report comparable results. In yet *another* comparable study, Santurkar *et al.* (2023) find poor simulation accuracy of political opinions when LLMs are asked to simulate 60 different US minority groups. Greater fine-tuning—what Santurkar *et al.* (2023, p. 29971) call "steering"—fails to improve accuracy.

For Santurkar *et al.* (2023), a major reason for poor simulation is poor training data. They suggest that popular LLMs like GPT-4 have simply not been trained on enough data

[114] By way of a primer, the typical silicon sampling study will consist of two components. Firstly, samples will be generated by prompting a model like GPT-4 to generate a synthetic profile for an individual. Individual characteristics could often be generated through random sampling of a representative data set, such as a national census. This will produce a sample of synthetic participants who match a representative sample of the target population, in terms of average demographics (or other data points), though no synthetic participant might perfectly match a 'real' person within the target population. Matching is often referred to as 'digital cloning' or just 'cloning.' It, too, is being tested as a behavioural research method, which will be discussed in more detail shortly (e.g., Park *et al.,* 2024).

Secondly, studies require behavioural data from the target population to compare the with 'behaviour' of the silicon sample. Without a comparison to a comparable group, the insights derived from the silicon sample are limited. Comparison is needed for initial verification of the simulation quality of the silicon sample. If verified, the sample may then be used to predict behaviour change within the target population without necessarily observing what the target population does. Though, to my knowledge, no silicon sampling study has examined the long-run accuracy of a silicon sample compared to a target group, to either verify predictions or monitor the longevity of the silicon sample's predictive powers.

originating from or representative of minority groups. A similar argument is given by Shrestha *et al.* (2024). In their study of policy opinions, Shrestha *et al.* (2024) report highly accurate simulations of opinions when simulating people from WEIRD countries—*western, education, industrialised, rich, and developed* countries—but poor accuracy when simulating those from non-WEIRD countries. On the one hand, such findings around minorities may be because these groups will naturally have smaller representation within a representative dataset. One study which might support this perspective comes from Gmyrek *et al.* (2024). Rather than examining political preferences, this study used GPT-4 and silicon sampling to simulate opinions about various occupations, such as prestige and perceived social value. Gmyrek *et al.* (2024) report that this simulation demonstrated high accuracy when occupations were grouped into high-level occupation categories (e.g., doctor), but less accuracy as more specific occupations were examined (e.g., oncologist). By virtue of being highly specific, details around specific occupations are likely to be less-represented within the training data of a model like GPT-4, leading to a drop in predictive accuracy.

On the other hand, under-representation in data may reflect an array of social and economic barriers to representation, which points to a more substantial methodological challenge for silicon sampling approaches (Sorensen *et al.*, 2024).[115] In such instances, under-representation reflects a *bias* against some groups, rather than merely *reflecting* their minority within society as a minority of examples within the dataset. This is a perspective which Argyle *et al.* (2023) acknowledge and believe is essential when evaluating algorithmic fidelity. Even if an LLM has "many biases," one bias may still exert an outsized influence on the final output, essentially *crowding out* other biases within the model (Sorensen *et al.*, 2024). For instance, Peterson (2024) has demonstrated that the training of LLMs typically involves the training 'long tails' *out* of the model. These long tails capture observations which deviate from the average, and which in social terms will often represent minority

[115] For instance, it made be difficult to train a model on data from countries where digital infrastructure is not common or well-developed. Furthermore, treating these data appropriately so as to retain meaningful insights requires programmers to have knowledge of and sensitivity to minority experiences, as captured in data. For instance, distinguishing colloquialisms and esoteric language from errors or some other anomaly which might be destroyed in data cleaning. One interesting case comes from the Japanese Government's own AI strategy, which notes that popular LLMs like GPT-3.5 are not trained on enough Japanese data to achieve a high quality for uses in Japanese society. This creates a political dilemma—should the Japanese Government take steps to enhance data sharing with a private, American company like OpenAI (creators of GPT-3.5), perhaps receiving more immediate benefits? Or, should focus be given on developing a Japanese-centric LLM, likely domestically, to ensure high quality within the Japan, and appropriate treatment of Japanese data?

groups and their views.[116] Removing these outliers causes LLMs to improve in terms of average performance, but at the expense of the average becoming *over-represented* within the model. Peterson (2024) thus argues that LLMs and similar AI systems are effective aggregators of populations but may promote an artificial consensus when deployed to *simulate* populations.[117]

Amirova *et al.* (2023) offer an additional, and very interesting, critique of silicon sampling. Focusing on the applicability of silicon sampling for *qualitative* research, Amirova *et al.* (2023) used GPT-3.5 to generate synthetic interviewees. The qualitative data which is subsequently generated was then analysed using qualitative research methods, rather than simply comparing quantitative measures (as many other silicon sampling studies do). Amirova *et al.* (2023, p. 1) find that in terms of key themes and broad topics, LLM simulations are "strikingly similar" to responses given by people whom they also interview. However, when a more detailed analysis of the interviews is undertaken, synthetic interviewees differ substantially from their human counterparts, including in terms of tone, structure, and language style. Amirova *et al.* (2023) thus conclude that silicon sampling fails to demonstrate anything more than a high-level approximation of human populations.

This is a potentially damning conclusion. For behavioural science, as for social science broadly, one of the major appeals of silicon sampling is presumably, to gain access to insights which are not immediately accessible. For instance, when there are limited qualified interviewees; when one needs insights much faster than is feasible. If silicon sampling succeeds primarily in simulating the most accessible group—the *average* person—then the benefits of such methods may be illusory. Agnew *et al.* (2024) build on this critique, and in the process, present a critical assessment of silicon sampling. In their review, Agnew *et al.* (2024) note that increased speed of data collection, followed by cost savings, are the most cited benefits of silicon sampling amongst studies exploring the approach. In their review, only around half of studies point to a greater diversity of perspectives as an advantage of silicon sampling.

Thus, with a critical eye, it is difficult to suggest that silicon sampling is at present a reliable method within behavioural science,

[116] 'Minority' here refers to any group which may be in the minority, depending on how one stratifies the population. Minorities and majorities may be constructed in numerous ways.

[117] One might contrast this argument with the use of word-embeddings to identify biases within populations. Bias detection is often concerned with aggregate behaviours, even despite behavioural science's increasing interest in personalisation. Hence why many biases are described as 'the *tendency* to do *x*.' This is to say, *on average*, a person will do *x*.

or the wider social sciences.[118] This is to be expected with any novel method; where important challenges around representation, measurement of accuracy, and practitioner motivation demand resolution. An alternative simulation method, utilising generative AI, may resolve some of these challenges associated with silicon sampling, though introduce others in turn.

Digital cloning is a broad term to describe the simulation of people using generative AI technologies. Digital clones can be created in several ways, and deployed for a multitude of ends, some more controversial than others. For instance, demographic and psychological data could be used to create a digital 'cognitive' clone of an individual. This person could then use the clone as a cognitive aid, helping them make decisions in everyday life. For instance, Golovianko *et al.* (2023) argue that critical decision-makers (e.g., doctors, executives, military commanders), whose absence may have an outsized impact on the outcomes of a decision, may utilise digital clones as 'donors' when they are not available (e.g., due to sickness).[119] This builds from previous work by Golovianko *et al.* (2021), who suggest digital 'physical' clones could act as avatars for a person, allowing them to be in many different places at once.[120] Digital cloning has even sparked discussion around *resurrection*, in a manner of speaking. If AI technologies can accurately simulate the mannerisms and personality of individuals, then one might choose to digitally clone a deceased loved one (Iwasaki, 2023); desires for

[118] Though, the same critical eye may conclude that the cost savings of silicon sampling, coupled with the *appearance* of promoting inclusivity, may motivate further experimentation and deployment of the approach in the coming years (Mills and Sætra, 2025).

[119] Golovianko *et al.* (2023) seem to focus on what in economics may be referred as to people with high 'human capital,' or in transaction cost economics specifically, as 'human specificity.' The latter is most relevant for this footnote. High human specificity means having specific knowledge and skills which demonstrate an outsized increase in value in a specific situation (e.g., Williamson, 1981). An oncologist, for instance, will have intimate knowledge of a cancer patient they have treated for several months; knowledge which will be most valuable *in relation to that specific patient*, and which will be difficult to replicate (not only because another doctor might not share a cancer specialism, but because other doctors will not have the social relationship with the patient that their long-term doctor would have). Similar arguments could be made for the executive of a large company, a general commanding an army in the field, or any other situation where a person contributes outsized value in a specific domain.

Noting this is important because it demonstrates an important economic criticism of a simulation methods like digital cloning. Transaction cost economists argue that high human specificity is a boon for those who have it, as it allows them to demand greater compensation (as their absence is more impactful, and their skills harder to re-acquire). Digital cloning may thus be understood as a kind of *automation*, which might have a downwards pressure on earnings (while having an upwards effect on productivity—this raises an interesting, but tangential, question about *who owns the digital clone*?).

[120] Note that a 'physical' digital clone would entail quite different data than a 'cognitive' clone and is also likely to have a different degree of autonomy. A 'physical' clone may be more like a tool, whereas a 'cognitive' clone might be used as a tool *and* as a machine.

a digital 'after life' might even prompt some to digitally clone themselves *while they are alive* as a kind of autobiographical (or self-obsessed) exercise. Applications such as these remain controversial (Iwasaki, 2023).

Of course, digital clones might be used for more conventional objectives. Truby and Brown (2021) argue that digital cloning is a natural extension of digital surveillance and micro-targeting practices which characterise modern online advertising and retailing. They suggest that firms may develop digital clones of people without meaningful consent,[121] using these clones to experiment with different marketing and persuasion strategies to more effectively advertise to 'real' consumers. In many ways, such a hypothesis aligns both with discussions of personalisation in Chapter 2, and forthcoming in Chapter 4, as well as discussions of persuasive AI chatbots found in this chapter.

Though, digital cloning—and specifically *cognitive* cloning—may achieve more benevolent ends when applied to behavioural science. Much of this work has been pioneered in three papers by researchers at Stanford and Google (Park *et al.*, 2024; Park *et al.,* 2023; Park *et al.,* 2022). One issue with simulations of any kind is achieving a large enough scale of simulation to model, in any meaningful way, real-world behaviours and activities.

In the first paper of the trilogy, Park *et al.* (2022) show that LLMs can be used to generate thousands of artificial agents, each with bespoke goals, ambitions, interests, and relationships. They then demonstrate that simulations involving these agents exhibit behaviour which humans cannot accurately distinguish from real-world community behaviour. This is *not* digital cloning, and in fact builds from the principle of algorithmic fidelity. As Park *et al.* (2022, p. 1) state, "[artificial agent] techniques are enabled by the observation that large language models' training data already includes a wide variety of positive and negative behavior[s]." Nevertheless, such a study is an important first step in building towards digital cloning and simulation by demonstrating that agents generated by AI can demonstrate human-led interactions.

The sequel comes from Park *et al.* (2023). As previously, this is not a digital cloning study. Instead, this study leverages generative AI even more to simulate much more complicated agent behaviour. 25 artificial agents are created and placed in an interactive sandbox to interact with one another. These agents have the same complex set of bespoke traits which define their backgrounds. However, their interactions are simulated

[121] By which I mean with only *nominal* consent; consent in the same way that one 'consents' to terms and conditions documents which they are never actually expected to read.

generatively through a LLM, with these interactions *changing* the artificial agents over time. In this sense, the agents *remember* and may simulate *planning* of their behaviours, becoming "*generative* agents" (Park *et al.*, 2023, p. 1, emphasis added). In one fascinating instance, a single agent proposed throwing a Valentine's Day party; subsequent agents then began to plan for this party, both by scheduling the party and finding fellow agents to bring to the party as dates. This leads the researchers to argue that the behaviour demonstrated in the simulation is meaningfully similar to human group behaviour, and much more complicated than simulation techniques lacking the generative AI component.

The final study of three introduces digital cloning. Having demonstrated that LLMs can generate agent profiles which support believable behavioural simulations, and that LLMs can allow agents to change, remember, and plan over time, Park *et al.* (2024) investigate how closely digital clones can simulate the behaviour of those they are cloned from. The authors interviewed 1,052 individuals about various aspects of their lives. A LLM was then used to create digital clones of these individuals from these interviews. Park *et al.* (2024) also collected data on social attitudes, personality data, and data about economic decision-making from the participants. Two weeks later, participants were invited back to retake these tests. Retesting allowed the researchers to compare participant accuracy to that of their digital clones, in a manner which essentially asks *how well do you know yourself?*

Compared to the original social attitudes survey, digital clones accurately predicted the attitudes of their 'counterparts' in 69% of instances. However, accounting for the rate at which participants themselves were consistent (around 81% consistency over two weeks), the normalised accuracy of the digital clones rises to around 85%. Correlations between clones and participants, in terms of personality and economic behaviour, were also relatively high at around 75%, and 60%, respectively. And, in terms of representativeness, Park *et al.* (2024) report no meaningful difference in accuracy for minority participants.[122] Though, such levels of accuracy were only found when clones were created from extensive interview data. Clones created using a simple personal description, or demographic data, did not demonstrate as high accuracy when simulating social attitudes (around 70% normalised accuracy). Interestingly, though, interviews did not outperform

[122] One might therefore speculate that some of the poor representation found in silicon sampling studies arises from 'participants' in those samples not being modelled on actual people. Representativeness is often much more than a statistical benchmark.

these 'persona' and 'demographic' clones when predicting personality or economic decision-making.[123]

The work of Park and colleagues is fascinating. Digital cloning and simulation may allow behavioural scientists to predict behaviours, both individually and as part of a group, which could offer invaluable insights. Furthermore, one can readily imagine experimenting with policy ideas in the simulation sandbox, observing how people are likely to respond to an intervention without having to *actually* experiment on people. This may reduce the risk of harm and save money (in some instances). Combining, say, Aonghusa and Michie's (2020) work on predicting policy recommendations with digital cloning may allow rapid *testing* of those predictions, and ultimately lead to better policy.[124]

Yet, one application of digital clones within behavioural science may stand out: organisational behaviour. Organisational behaviour has been little discussed in this chapter, in part because modern behavioural science tends to focus on individual behaviour (e.g., biases) and outcomes (e.g., discrete choices).[125] Thus, the recommendations which modern behavioural science prescribes (e.g., nudging) may gain less purchase within a complicated, organisational context. Though, organisations are increasingly entering into behavioural science discussions as practitioners seek both to contribute to more challenging, complex behavioural problems (Hallsworth, 2023; Sanders, Snijders and Hallsworth, 2018), and as pressure pushes the field to demonstrate relevance in increasingly challenging, complex policy areas (Chater and Loewenstein, 2022).[126]

[123] An immediate implication of this is that a person's personality may be inferred from a digital clone created using only an assortment of demographic data, provided an appropriate generative AI infrastructure for simulation is available. Furthermore, personal data—collected through a quasi-interview—may improve digital clones. Such observations should return one's thinking to prior discussions of persuasive AI chatbots.

[124] This is not to overlook the question of *who owns digital clones* and *who controls them*? Assuming such clones are powerful predictive devices, these questions gain immediate importance. One should also not overlook the risk that a promise of superior prediction paves the way for an effective surveillance state (much as the promise of targeted advertising and recommendation has allowed a private sector surveillance apparatus to flourish). The ideas described in this chapter could easily lead one to imagine a world where interactions with AI chatbots, as one searches for necessary government documents, are used to create an accurate digital clone of a citizen, a clone which is then endlessly experimented on to predict how the same government should treat that citizen.

[125] This is not to suggest that *the behavioural sciences* have neglected organisations. Organisational psychologists, social psychologists, organisational studies scholars, managerial decision-making scholars, and so on, would all protest the notion that they overlook organisations as a unit of behavioural analysis, and they would be right to do so.

[126] There is another, tangential reason, to care about organisations within a discussion of AI applications to behavioural science. This is because, in a manner

Hallsworth (2023, p. 312) has called for behavioural science to "see the system" emphasising that behavioural science interventions do not happen in a vacuum, but instead within a context which produces ripple effects and feedback mechanisms. This call is combined with a call to bring insights from systems thinking and systems analysis into behavioural science practice.[127] One attempt to do this might be to target interventions based on their place within a hierarchy, or within a social network. For instance, influencing the behaviour of a manager might have a greater overall effect than influencing a worker, as the manager's behaviour is likely to have ripple effects impacting those who work under them. Understanding such dynamics has long been of interest to behavioural scientists (Dolan and Galizzi, 2015; Sanders, Snijders and Hallsworth, 2018).[128] One may speculate about the

of speaking, intelligent behaviour emerges from individuals *in the same way* as it emerges from groups of individuals; and, as it emerges from deep learning AI. This is to say, individual cognition, organisational decision-making, and AI prediction, may all be isomorphs of each other (Simon, 1981). This broad argument begins in Minsky's (1986) notion of the 'society of mind' and is elaborated upon by Turkle (1988). Simply, individual neurones in the human fire, and it is the aggregate effect of firing (strength and timing) which governs a person's behaviour. Likewise, nodes in the deep learning model 'light up' when inputs breach an activation point, with the aggregate 'brightness' of nodes determining the 'action' of the model. But interestingly, one might also understand the organisation in the same way, though rather than neurones or nodes, individuals and teams are the units which 'fire' or 'light up' to determine the overall behaviour of the organisation.

The foundation of this idea is provided in Cyert and March's (1992) behavioural theory of the firm, which in part emphasises how different stakeholders within a firm pursue their individual goals *through* the firm, with the firm's success determined by the extent to which the means for stakeholders to achieve their goals complement, rather than contradict, one another (also see Simon's *Administrative Behaviour*). This, according to Simon (1997a), is the essence of *cooperation*. Likewise, the extent to which a node 'lights up' is based, perhaps, on the extent to which nodes in the previous layer 'cooperate' and 'light up' at the same time. It is perhaps telling that Minsky (1986), in describing intelligence as a 'society' or collective of simpler information processing units, discusses democracy. Neither is it surprising that Hayek, whose *Sensory Order* (Hayek, 1999) deals with connectionist psychology, also championed the free market as an efficient, information processing unit made up of decentralised individuals each pursuing their own objectives (Hayek, 1945). This notion—that smaller, simpler information processing units may, in the aggregate, achieve successful, complex, *intelligent* behaviour—is an isomorphic description which connects brains, organisations, democracies, and now, machines.

This is not to suggest this isomorphism is *all there is* to the behaviour of these information processing systems. Rather, it is to point out that organisations have a place in this conversation, both as entities which exhibit intelligent behaviour and—insofar as artificial intelligence is simply that which is not natural (human) intelligence—as a kind of *AI* (Simon, 1981). This, more than any other, is the golden thread through Herbert Simon's eclectic research career. Davies' (2024) recent work on management cybernetics also makes the link between AI and organisation, though through Stafford Beer rather than Simon.

[127] Hence, again, why behavioural science should be understood as a relative of *cybernetics*.

[128] One name which appears to have not caught on is "network nudge" (Sanders, Snijders and Hallsworth, 2018, p. 160).

applications of AI to such problems. For instance, a predictive AI system (or even a relatively simple algorithm) might be able to score the 'influence' of each member of an organisation, so that interventions may be targeted at optimal leverage points. Following Chapter 2, AI systems might be used to personalise interventions not to the individual, but to the *organisation*, insofar as such personalisation is done to manage the spread of the intervention (or the intervention's effect) throughout the organisation.

One challenge with understanding organisations, and thus successfully intervening to affect change within them, is that an organisation rarely behaves in a manner congruent with how it is, on paper, organised (Peters, 2017; Simon, 1997a). One might be able to sketch out the hierarchy of an organisation and rank the 'influence' of its members by each's position within that hierarchy.[129] One might use some form of network analysis, such as looking at email correspondence, to approximate who contributes most to driving the dynamics of an organisation. But such perspectives ignore important, often undetectable interpersonal factors which might, across time and hierarchy, matter much more to the behaviour of an organisation, and its members (Simon, 1997a).

Imagine a simple organisational structure consisting of person A—a senior manager—B—a junior manager—and C—an entry-level worker. On paper, A > B > C in terms of organisational authority. If one wished to affect the behaviour of B and C, targeting A would seem to make a lot of sense in this *linear* hierarchy. But organisational scholars—if not behavioural scientists—will frequently stress that members of an organisation regularly embody *several different* roles at the same time (Simon, 1997a). A is not just a manager; they might be a sports fan. And C, while an entry-level worker, might support the same team as A. All

[129] The history of 'hierarchy' is fascinating, and perhaps explains why this word continues to be one which is disliked by many people. Peters (2017) notes that the original use of the term was in association with the Church, arguably the first major organisation in Europe. The Church's power allowed it to dominate the lives of peasants (and, often, the nobility, too) and extract resources through tithes. Equally, the Church often legitimised rulers who would also dominate the masses, extract taxes, and so on. Both likely contributed to people souring on the idea of hierarchy. It likely did not help, either, that the notion of hierarchy would be applied by Dante to his description of Hell, a literary trope which can be found in, say, Kafka's critique of bureaucracy (see Graeber, 2015).

For the interested reader, one can carry this critique through to understand discussions of another word often misunderstood today—*the state*. When Smith, Hume, and Locke were attacking the 'state' and advocating for a new, liberal arrangement of society, they were not attacking *today's* democratic state tasked with provision of public services and international cooperation. They were attacking *absolute monarchies*, and by extension, organisations such as the Church which were integral to the monarch's power to the delaying of liberal reforms. Nevertheless, this historical context is often lost in modern discussions by 'classical liberals' leaving their critiques to fall into a Kuhnian trap.

the while, B supports a bitter rival. At the coffee machine or the water cooler each morning, then, one might often observe a situation where A = C, and even where C > A. These competing maps—or *graphs*—of the organisation exist simultaneously (e.g., A > B > C *and* C >= A > B), and lead to a fundamental challenge around the question of influence—should an intervention target A, or C?[130]

Peters (2017) resurrects a cybernetic term—*heterarchy*—to describe such problems. Heterarchy was originally proposed by Warren McCulloch (1943)—another 'father' of AI—to describe how complex behaviour arises within systems of agents. One can imagine that in an organisation described as A > B > C, complex behaviour cannot emerge. A is always the focal point; B responds to A's actions, and C to B's. Hence why, as above, this hierarchy can be described as linear. But in the second description, where C >= A > B *and* A > B > C, much more complicated behaviours can arise. For instance, B can direct C to perform a task; C can then perform the task, or attempt to influence A, who can then direct B to do something different, which then impacts C, and so on.[131] Heterarchy, then, is a view of system behaviour which incorporates both the formal and informal connections between agents within the system, to use Simon's (1997a) language,[132] and holds that systems "are neither ordered nor disordered but instead are ordered complexly in ways that cannot be described linearly" (Peters, 2017, p. 23).[133] And, as Peters (2017) argues, it is often the

[130] On the matter of authority in organisations, Simon (1997a, p. 198) writes that the "formal scheme of organization will always differ from the organization as it actually operates in several important respects. First, there will be many omissions in it—the actual organization will exhibit many interpersonal relationships that are nowhere specified in the formal scheme… Second, the interpersonal relations in the organization as it operates may be in actual contradiction to the specifications."

[131] From this description, one sees that the heterarchical view incorporates feedback loops into organisational dynamics. For instance, in the described scenario, C *indirectly* affects their own behaviour, by influencing the behaviour of B through A.

[132] Simon (1997a, p. 198): "The term "informal organization" refers to interpersonal relations in the organization that affect decisions within it but either are omitted from the formal scheme or are not consistent with that scheme. It would probably be fair to say that no formal organization will operate effectively without an accompanying informal organization."

[133] One (say, an economist) might object that while such social connections are not unrealistic, A, B, and C remain ultimately bound by the authority of the organisation, and that an embrace of the informal may be mitigated through adequate organisational incentives (e.g., Jensen and Meckling, 1976). Such an objection is why it is important to incorporate bounded rationality theories of organisations (e.g., Cyert and March, 1992; Simon, 1997a) into the analysis, as these perspectives emphasise—both theoretically and through observation—that sensitivity to incentives is often less than natural propensity to messy, but *humans*, interpersonal dynamics.

Furthermore, as Simon (1997a, p. 183) discusses, authority or influence are rarely shown through punishments or sanctions on disobedience: "Any study 'informal' structure of a system which determines how it changes, and thus how the system *can be changed*.[134]

Behavioural scientists who wish to 'see the system' might be said to have embraced the idea of heterarchy insofar as there is growing discussion of 'behavioural systems mapping.' Behavioural

of power relations which confines itself to instances where the sanctions of power were invoked misses the essential fact of the situation. To avoid this fallacy, authority [must be] defined… not in terms of sanctions of the superior, but in terms of the behaviors of the subordinate." To this end, Simon (1997a, p. 177, original emphasis) emphasises that, "it is necessary to keep constantly in mind the idea of a decision as a conclusion drawn from a set of premises—value premises and factual premises. Organizational influence upon the individual may then be interpreted not as determination by the organization of the decisions of the individual, but as determination for him of *some* of the premises upon which his decisions are based." Thus, authority is much more about telling someone *how* something should be done, rather than *what* is to be done. The former is more susceptible to interpersonal informality.

[134] Peters (2017) uses the idea of heterarchy to describe the puzzle of why the Soviet Union never succeeded in developing their own version of the internet. The Soviet economy was built around the idea of central planning, which naturally leant itself to information and communication networking. The Soviet Union also did not lack for scientific talent, technological vision, or raw materials. Soviet pioneers of a Soviet internet, as Peters notes, also regularly had high-level political support. Yet, Peters argues internet projects failed in the Soviet Union because they were designed to work within the planned economy *as it existed on paper*. In reality, the Soviet Economy ran off an extensive black market and informal economic arrangements. Not only would it have been extremely difficult to design a communications system which accounted for this enhanced complexity, it would have also put these designers in the crosshairs both of those who benefited from the underground economy, and those who could not allow government acknowledgement and endorsement of it.

While Peters highlights aspects of the Soviet economy which undermined that nation's communication objectives, it is important to appreciate that heterarchy is not a way of describing the failures of a centralised hierarchy. All organisations can be understood as heterarchies, insofar as one recognises that members of an organisation are not bound solely by their position within the formal organisational structure. For instance, it is often perplexing to scientists why their policy ideas get rejected, despite ample evidence supporting their advocacy. Yet, Kingdon's (2003) fascinating work on government agenda setting shows that ideas live and die not by their merits *per se* (though this is important), but by a complexity of social and political conditions. In articulating this point, Kingdon discusses the 'garbage can' model of decision-making (Cohen, March and Olsen, 1972) which suggests that rather than decisions being made through a rational analysis, ideas are thrown into a metaphorical garbage can, and *then* problems are found to which the ideas can be applied. Both Kingdon's work, and the garbage can model, point to a more heterarchical view of organisations (note that James March, of the garbage can model, was also a frequent collaborator of Herbert Simon, and was one of half Cyert and March's *behavioural theory of the firm*).

Does this mean that the formal description of an organisation is useless? To this question, Simon (1997a, p. 198) retorts that the formal structure serves two important purposes. Firstly, it (should) set "limits to the informal relations that are permitted to develop within it." Secondly, it (should) "structure [the organization] to encourage the development of the [informal relations] along constructive lines." That the informal structure of the Soviet economy derailed internet projects *despite* support within the formal structure is indicative of a formal structure which failed to either control or shape inevitable informal relations.

systems mapping seeks to draw a 'map' of "the key actors in a scenario and their behaviours" and "how these behaviours relate to each other" (West *et al.*, 2020, p. 27). In the above hypothetical, the notions of A > B > C and C >= A > B could be viewed as rudimentary system maps, insofar as one can identify 'nodes' (A, B, and C) and draw arrows of influence connecting these nodes. One potential difficulty of a behavioural systems mapping approach, however, is that it is not clear how different maps can exist *simultaneously*, and thus capture a multitude of different roles which individuals of an organisation may adopt.[135] For instance, what if, for some tasks, A is willing to overrule B's command of C, but in other instances, A steadfastly backs B? Which map should be used, and when? As Peters (2017, p. 23), writing on McCulloch, notes, "even the simplest systems can be subject to multiple competing regimes of evaluation."

Simulations with digital clones may resolve these problems and allow behavioural science and behavioural systems mapping to incorporate heterarchical thinking into the design and implementation of interventions. This is not to suggest that a single simulation predicts the future of a fundamentally complex (if not chaotic) system. Rather, it is to suggest that through multiple simulations of digital clones interacting with one another, one might be able to observe patterns and trends in those interactions which reflect the heterarchical structure of a real organisation; a structure which may be difficult to observe through other methods, which may elude even the organisation itself; and if not, a structure which the organisation might never willingly admit is demonstrative of how it operates.

Solutions With Problems

This chapter has reviewed various applications of AI within behavioural science. To conclude this chapter, it is naïve to suggest that AI will not impact behavioural science. Equally, it is naïve to suggest that behavioural science will necessarily undergo a radical transformation. Substantial doubt persists around silicon sampling, while evidence supporting AI applications remains sparse in several instances.

The focus of this chapter has been on how AI can support behavioural science to tackle the problems of too much information, and obscured or hidden information. Though, from the outset, the reader has been encouraged to understand information not in terms of *bits*, but as something with a social life. In doing so, rather than merely describing applications, this chapter

[135] The exhibition of a role, under what conditions, and so on, may also be subject to substantial randomness. This is but one random variable which might influence how organisations behave.

has considered how similar applications could be applied to different *ends*, and how the merits of various applications become complicated when considered within different social contexts. The following chapter, Chapter 4, elaborates on many of the social implications of AI in behavioural science which this chapter has begun to unpick.

Chapter [4]—Algorithms and Their Discontents

> "[I]f instead of focusing on what machines can't do, you grant, at least for the sake of argument, the idea of a machine capable of meeting any specifications, you can still ask, *So what?*'
>
> —Sherry Turkle, The Second Self (2004, p. 241, emphasis added)

UI, not AI

Discussions of technology can sometimes fail to adequately distinguish between three albeit related categories: how technology *can* be used, how it *ought* to be used, and how it *actually is* used. The previous chapters have largely dealt with the first category. This chapter, and Chapter 5, will focus more on the latter two.

In the 1960s, the philosopher and computer scientist Joseph Weizenbaum developed ELIZA, a computer programme which today would be considered a chatbot. ELIZA was simple in comparison to modern LLMs, though apparently remarkable for its time. Participants could communicate with ELIZA through a typewriter connected to a computer running the programme, and ELIZA would produce natural language responses. Weizenbaum discovered that people found ELIZA very engaging and had enduring conversations with the programme. The programme itself became something of a celebrity on the MIT campus where Weizenbaum worked and was sufficiently compelling to secure its spot in the history of computing and AI (Turkle, 2004; Weizenbaum, 1976).

One of the reasons Weizenbaum (1976) developed ELIZA was to demonstrate to a non-technical audience the advances in computing power in only a short period of time. He noted that computers, then as now, were regarded as difficult and non-intuitive devices. People struggled to understand quite *what* computers did, from a technical perspective. Trying to convey to a wide audience how computers were advancing was thus a challenge, and an important one given how Weizenbaum and others anticipated computers impacting society.[136] ELIZA was a

[136] For avoidance of doubt, Weizenbaum was often critical of computer applications and AI evangelists. His experiences with ELIZA, which will be further discussed in this chapter, helped mould several of his views about the limits of computers within human affairs. Thus, the above phrasing is not to suggest that Weizenbaum felt computers *ought* to change society, but rather, that he simply recognised they would.

response to this challenge. By creating an intuitive interface, the perceived technical barriers to using computers were reduced. Rather than having to code, one simply needed to write, as if one were talking to a friend or colleague. Equally, ELIZA simulated an activity—a conversation—which people undertook every day. One did not need to know *what* the programme was doing to appreciate that it was doing something requiring advanced technology. Such framing, Weizenbaum hoped, would encourage many more people to engage in discussions about advancing computer technologies—*if computers can talk, imagine what they will do next?*

ELIZA offered people an accessible way of experiencing computers, creating the opportunity for computer scientists to engage a wider audience in discussions about the new advances in the field. ELIZA exploited the rapidly growing processing power of computers, but it was not in itself an incredible new development of computational (or artificial) intelligence; only a "small and simple step" in the development of natural language processing, *if that* (Weizenbaum, 1976, p. 7).[137]

In November 2022, ChatGPT was released. For perhaps a week, I took no notice of it. I had seen a flurry of social media posts containing screenshots of conversations with the application, which was using a more advanced version of GPT-3 to generate natural language responses to text inputs. My apathy, or perhaps *ignorance*, came because I had seen GPT-3 in action before. For several years, several AI models, including GPT-3, had been available for anyone to use via an application programming interface, or API. With some knowledge of programming and a developer account with OpenAI, the creator of GPT-3, one could use the model to generate text.[138] As a result, all the screenshots of ChatGPT responses failed to resonate with me—I had already seen such text generation.

Then things clicked. ChatGPT differed in what I had seen before in two ways. Firstly, there was the genuine development in the underlying model, with GPT-3.5 being a more advanced version of GPT-3. But secondly, and more importantly, there was the user interface itself. In a manner reminiscent of ELIZA, ChatGPT was not a huge technological leap, but rather, a reframing of an existing technology to make said technology more accessible to a non-technical audience. By allowing people to interact with GPT-3.5 through a search bar, millions of people could suddenly begin using AI technologies without having an above-average level

[137] I will return to what ELIZA was *actually* doing, and how ELIZA *actually* worked, later in this chapter.

[138] My first interaction was even simpler, using a readily downable model called GPT-Neo, based on GPT-3's precursor, GPT-2. This still required programming knowledge but avoided APIs.

of programming knowledge.[139] For many people, text generation of the sophistication of GPT-3 *was* novel, and was worth sharing.[140]

In both the ELIZA story, and ChatGPT, one sees the importance of the *user interface*, or UI. In this chapter, UI can be thought of in two ways. Firstly, there is the literal UI—the interface through which humans interact with AI. Secondly, there is the behavioural UI—how a person perceives, understands, and subsequently uses AI systems. These two perspectives on UI are connected and cannot always be neatly separated.[141] A reader is asked to simply keep these categories of UI in mind throughout this chapter, both to draw one's own links to these ideas, and to aid in clarification of links where they may be found.

*

A prominent theme in recent discussions of AI applications in behavioural science is how AI can be used to 'debias' people, or in other words, to help people make better decisions by encouraging them to choose different outcomes (Sunstein, 2024). This idea is quite closely linked to arguments surrounding choice engines and personalised paternalism discussed in Chapter 2. There are also pronounced links to discussions of professional decision-makers, such as doctors and judges, as discussed in Chapter 3. The crux of this argument is that people are biased in a myriad of ways, from

[139] Note that 'above-average' here does not mean an especially high level of knowledge. When one considers that most people do not know *any* programming, even a small amount of programming—and certainly enough to use GPT-3 or GPT-Neo—is 'above-average.'

[140] An important difference between these stories, in terms of bringing technology to the masses, is the motivation of Weizenbaum compared to OpenAI. Weizenbaum wanted to share developments in computer science and facilitate discussions with non-technical people who could contribute to ideas about how computers could be used, and *ought to be* used. While speculation, it is not unreasonable to suggest that OpenAI's launch of ChatGPT was driven by economic motives. LLMs are enormously expensive to develop and run, and it is likely that OpenAI recognised that further development of their models required huge sums of additional investment. Showcasing their technology, and building a large user-base, was likely essential in supporting the company's investment ambitions (Mills, 2024b).

[141] Though, as an astute reader might guess from the preamble in this chapter, this is hardly a novel argument. Weizenbaum's experiments with ELIZA taught him much about the strange ways in which people *experience*, and thus *understand*, technology. Turkle's (2004) *The Second Self*, which explores much of Weizenbaum's work up to that point (1984), is also a detailed exploration of how technology and AI changes how people understand themselves. Both could be linked to Heidegger's (2010) arguments about how tools transform us as much as we use tools to transform the world, an argument made more implicitly by Illich (1973), too, or Marcuse's (2013) *One Dimensional Man*. More recent efforts include Negroponte's (1995) *Being Digital*, Scott's (2015) *The Four-Dimensional Man*, Lawrence's (2024) *The Atomic Human*, and Turkle's (2013) *Alone Together.*

overvaluing the status quo (Samuelson and Zeckhauser, 1988) to relying too much on similarity, salience, and those ideas most readily brought to mind (Tversky and Kahneman, 1974).[142] But algorithms can help avoid these biases by standardising decision-making processes (Sunstein, 2023, 2022b). The widespread use of algorithms in society, according to Sunstein, would lead to more equitable social outcomes and more efficient allocation of societal resources precisely by removing the biased discretion of professional decision-makers.[143]

Such a proposal is contentious. One controversy may be that the notion of algorithms *counteracting* biases runs counter to the headline arguments of those in the algorithmic bias literature that algorithmic systems, from simple computers to complicated AI systems, often contain, recreate, and perpetuate biases found in society (Kordzadeh and Ghasemaghaei, 2022). This has already been seen in discussions of word embedding models which encode racial and gender biases in Chapter 3 (Bolukbasi *et al.*, 2016). One approach might be to distinguish between discriminatory biases (d-biases), and cognitive biases (c-biases). The former are systematic biases, but for prejudicial reasons which new information would not overcome. The latter are systematic also, but arise for reasons of excess information, poor information, limited cognitive power, or inhibitive environments to deploy cognitive power. Such a distinction is important, as much of the algorithmic bias literature deals with d-biases, while the behavioural science literature aspires to tackle c-biases. Insofar as one focuses on c-biases, it is a reasonable hypothesis that algorithms, with greater information processing capabilities, would be less c-biased than a person, and thus work as a means of debiasing decisions (Mills, Costa and Sunstein, 2023; Sunstein, 2022b).[144]

[142] People are also regularly found to think about risk and uncertainty in ways which do not align with the mathematics of these domains, leading to poor decision-making in areas of, say, gambling or investing (Camerer, 1989, 1987)—and, by extension, in healthcare, criminal justice, and more (Sunstein, 2023).

[143] Writing on the use of algorithms in medical settings, Greene and Lea (2019) offer a similar argument to that above. They note that advances in medicine now mean that doctors face more complicated decision-making than in previous decades, necessitating the effective use of significantly more data than a person can reasonably handle. To this end, Greene and Lea suggest that algorithms and AI systems may be useful aids for doctors. As this chapter will explore, their may be important differences between the use of an algorithm because of a changing decision-making environment, as Greene and Lea (and others) suggest, and algorithmic deployment because of *inherent* human biases.

[144] This is not to dismiss the importance of d-biases, and it is essential to appreciate that conceptual distinctions do not always map onto actual discussions. For instance, mugshot bias—the overweighing of a defendant's mugshot in a bail hearing—is *clearly* a d-bias, even if implicitly. Nevertheless, Sunstein (2023) discusses mugshot bias is the same way as recent offence bias—the overweighing of the most recent offence relative to the whole offending

Much has been said on the topic of algorithmic bias (e.g., Kordzadeh and Ghasemaghaei, 2022). This chapter will leave many of these interesting discussions for others and concern itself centrally with complaints which might emerge from these recent (but not necessarily *new*) claims of algorithms tackling c-biases. Two complaints are relevant.

The first is what one might call the *value-bias problem* or the *so what?* argument. Broadly, it argues that evidence of c-biased decisions is not in itself justification for the use of algorithms to assist decision-makers. What is labelled as a bias is the product of value judgements about what is right, and what is wrong, what is fair, and what is not, and so on. Evidence of algorithm usage in public life suggests, contrary to intuition, that sometimes 'biased' behaviour is much more defensible than 'de-biased' behaviour. To an extent, calls for algorithms and AI to only nudge decision-makers, rather than dictate actions, may reflect this need for discretion (Sunstein, 2024, 2023, 2022b, 2019). Yet, these calls are often in response to the possibility that the *algorithm* gets something wrong—not that 'debiasing' is sometimes socially undesirable.

Secondly, algorithms and AI systems do not fundamentally change the social and institutional structures in which they are used. Indeed, if such technologies challenged these institutions, one might be sceptical to as whether they would be introduced at all. Algorithms might be used to support and perpetuate harmful behaviours, rather than challenge them. In some instances, AI systems might serve a *technosolutionist* function, supporting the continuance of inequitable and inefficient processes through what I call *machine laundering*.[145] These ideas are explored through an examination of the emerging literature on *selective adherence*.

The chapter ends with two sections that invite the reader to zoom out once more, and consider *how* AI systems should be used, given *what* AI systems can actually do *versus* what actually *needs to be done*. I explore how the decision to use algorithms and AI itself may be flawed, and that often such decisions must be taken within a wider social and organisational context. To this end, the chapter returns to the story of ELIZA, and of ChatGPT, and asks where people fit into things.

Another Bit in The Wall

The *value-bias* problem arises when social values can be interpreted or described in terms of cognitive biases. It thus concerns that of

history. This bias is much better described as a c-bias. For both, Sunstein (2023) advocates for the use of algorithms to support bail decisions.

[145] I cannot claim that this is an original term. I have definitely stolen it from someone else. Unfortunately, I cannot remember who I have stolen it from. To this end, I should be given no credit.

the normative status of biases. The idea of a cognitive bias is often framed as some objective phenomenon—as a systematic statistical deviation from some benchmark (Wilke and Mata, 2012). This masks the reality that biases are *always* normative determinations.[146] In some instances, this causes few issues. Consider once more mugshot bias—the finding that a defendant's mugshot significantly predicts the likelihood of a judge granting bail (Kleinberg *et al.*, 2018). Accepting that such a finding is accurate,[147] one may turn to the normative question: *should this statistical result be considered a bias worth mitigating?* Most people would agree that this result is a bias, and a bias which should be tackled through some change in the judicial process, potentially including (but certainly not limited to) the introduction of an AI system to independently evaluate (and thus nudge) a judge's decision. This is because mugshot bias is a d-bias, and most people abhor discriminatory practices within their communities. If polled, one would imagine a compelling majority in opposition to the practice of using a defendant's mugshot within a bail decision.[148]

The same normative defence of the mugshot bias cannot be levied at the current offence bias—the finding that a defendant's current offence significantly predicts the likelihood of receiving bail. For avoidance of doubt, the current offence bias is a c-bias; it is the result of too much information leading to the use of simplifying heuristics to eliminate information, and reach a conclusion (Sunstein, 2022b). This, though, does not protect current offence bias from the normative question: *should this statistical result be considered a bias worth mitigating?* Unlike the mugshot bias, one might suggest that the answer here is hardly clear cut.[149]

In the affirmative, one might consider the following scenario:

1) mom and pop, who own a mom-and-pop store, have never committed a violent crime, but they have evaded

[146] For instance, Dhami and Sunstein (2023) note that all cognitive biases in the Kahneman-Tversky research programme are essentially measurements of deviations from economic rationality. Of course, the selection of benchmark implies some normative judgement, and it is naïve to contend that the choice of benchmark does not reveal some preference, on the part of the benchmarker, for what behaviour *ought* to be. Also see Haselton *et al.* (2015) and Wilke and Mata (2012). Likewise, one cannot expect people to always separate scientific uses of words from their everyday usage and connotations. *Bias* suffers particularly in this regard.

[147] I see no reason to dispute this result, and generally do not intend this section, or chapter, or book to be a direct challenge to such findings.

[148] Note that this is not to suggest that judges *intentionally* discriminate on the basis of a defendant's mugshot—though some might. Rather, it is to recognise that even an implicit use of the mugshot is likely to result in discrimination by playing on stereotypes and cultural conditioning.

[149] An alternative way of phrasing the 'normative question' which might be more aspirational and thus appealing, is: *in the society we wish to live in, should this result be treated as a bias?*

taxes for several years, costing the government hundreds of thousands of dollars.

2) a career criminal, upon returning to their community, has stolen a smartphone, resulting in the loss of around one thousand dollars.

In this scenario, those who believe the current offence bias should be considered as such would point out that the career criminal's past crimes mean, statistically speaking, they are more likely to commit a crime if granted bail, compared to mom and pop. Yes, the former caused more *economic* damage, but their single offence implies a low propensity, and thus risk, to commit additional crime, while the *nature* of the crime does not imply mom and pop are likely to bring trauma or violence upon their community, if bailed.

In the negative, though, one might launch the following argument: 1) the criminal justice system is designed to first punish, then reform, and finally forgive those who commit crimes; 2) to judge someone for crimes for which they have already been punished, reformed, and forgiven, is to undermine the ethical principles of the criminal justice system; 3) if a person goes on to commit a crime after being released, and if their past crimes are highly predictive of future criminal activity, this implies that the criminal justice system is failing in its function to reform those who pass through it. Such an argument points out that current offence bias demands we bend some of our social ideals. In this instance, accepting that crimes for which one has been punished for should influence *future* judicial decisions—to accept that clean slates should not exist.[150]

The existence of a statistical artefact does not justify the imposition of an algorithm or AI system to act on (in this instance, to counteract) this artefact. In the case of the current offence bias, it is both possible that some judges ignore the criminal history because the current offence is most salient, *and* that some judges ignore the criminal history because they are tasked with forming a bail decision based on the principle that one should be judged only for their *current* crimes, not for those already served. To this end, an advocate for an AI advisor might suggest that the AI advice would be at the discretion of the judge to ignore—no proposal

[150] Furthermore, current offence bias may distract one from the actual solution to past crimes predicting future crimes—namely, criminal justice reform, investment in post-prison services, and revitalisation of communities through economic investment and social empowerment. These reforms are difficult, being economically costly and politically challenging. An algorithmic prediction model, in contrast, is easier—it is cheap and politically less taxing. This problem of easy *versus* hard decisions, or, more properly, individual *versus* institutional interventions, reoccurs throughout this chapter, as a reader will see. Also see Chater and Loewenstein (2023), Curchin (2017), Fuller (2020), and Mills (2024a).

today advocates for the *replacement* of judges with algorithms (Sunstein, 2024). This is not an unreasonable argument, but the normative component of biases is rarely the motivation for maintaining professional discretion. For instance, Sunstein (2024, 2023, 2022b, 2019) calls for professional discretion because algorithms do not always make accurate predictions. The need for discretion due to normative disagreements is generally absent from discussion.

This is a meaningful oversight given an increasing number of controversies which suggest the public care more about the maintenance of community values than they do about counteracting construed biases.[151] One interesting example to consider is that of a grading algorithm used by the UK Department for Education in 2020.

Of the numerous implications of the COVID-19 pandemic, that 16-, 17-, and 18-year-olds could not sit in-person examinations became a high-profile issue in the UK around May of 2020. With remote examination being fraught with the possibility of cheating, undermining the integrity of the exam system, the Department for Education, within the UK Government, announced that an algorithm would be used to assign hundreds of thousands of grades to students. This decision was defended on three fronts. Firstly, that few alternative options existed, given the pandemic.[152] Secondly, the alternative that did exist—the use of predicted grades given to each student by their teacher—was unlikely to produce accurate outcomes.[153] This is because teachers tend to suffer from what one might call a *grading bias*, a form of optimism bias whereby teachers predict higher grades than students actually achieve (Ofqual, 2020). Without recourse, predicted grades alone were

[151] There are a growing number of scandals which could be discussed here. These include the Dutch *toeslagenaffaire*, where an algorithm erroneously accused Dutch citizens of defrauding the child benefits system, and the Australian *robodebt* scandal, where an algorithm erroneously accused thousands of benefit recipients of receiving *too much* money, hounding them to return it. I have chosen to focus on the UK Department for Education's use of a grading algorithm during the COVID-19 pandemic as it is not so much a scandal resulting from an algorithmic *error* (as toeslagenaffaire and robodebt were), but one where the algorithmic solution was deemed less preferable than the bias it was designed to solve. The toeslagenaffaire and robodebt will be discussed, in a different context, later in this chapter.

[152] Ofqual (2020, para. 3), the body which oversees UK schools, claimed to have advised the Department of Education, "that the best option in terms of valid qualifications would be to hold exams in a socially distanced manner." Failing this, Ofqual advised the use of a standardisation model (an algorithm) to assign student grades.

[153] For avoidance of doubt, 'predicted grades' or 'predicted results' are common terms surrounding UK school exams. Every year, teachers are asked to predict individual results to give schools, students, and the government a ballpark estimate of likely performance. The use of an algorithm to *predict* and then *assign* grades to students was novel in the pandemic.

likely to offer an overly-optimistic estimate of student performance.[154] Thirdly, an algorithm could be used to adjust predicted grades to more accurately predict student performance. By combining historical trends data, previous student performance data, and school-level data with predicted grades, the Department of Education suggested algorithms could reduce the grading bias (Kolkman, 2020).

Ofqual (2020, para. 6), the body which oversees UK schools, stated that, "the principle of moderating teacher grades was accepted as a sound one." Yet, almost immediately following the release of the algorithmically assigned results, a huge public backlash began. The algorithm downgraded nearly 40% of those grades predicted by teachers, while only upgrading around 2%. While the majority of grades—nearly 59%—remained unchanged from those predicted by teachers,[155] the enormous downgrading prompted accusations that the algorithm was unfair. The public largely interpreted the perceived unfairness of the algorithm in two ways.

On the one hand, reports ran of superstar students from deprived areas now being downgraded because they were associated with an underperforming school. Here, there was an emphasis on children not being given the opportunity to demonstrate their individual talent. On the other hand, there was upset at the perceived bias in favour of those from more secure, economically prosperous backgrounds. Analysis seemed to support this—private, fee-paying schools saw the largest year-on-year grade increase, while publicly funded secondary schools and colleges saw the smallest increase (Nye and Thomson, 2020).[156] Private schools

[154] Ofqual (2020, paras. 6-7): "We were asked to implement a system of grading using standardised teacher assessments, and directed to ensure that any model did not lead to excessive grade inflation compared with last year's results… All the evidence shows that teachers vary considerably in the generosity of their grading… Using statistics to iron out these differences and ensure consistency looked, in principle, to be a good idea."

[155] How should one treat the 59% figure? One perspective is that it might be seen as a point in the algorithm's favour, suggesting an accuracy quite a bit higher than chance—assuming the algorithm tries to predict the grades given by teachers. Another might take an opposing view, though one still implicitly in favour of the algorithm: the huge downgrading is evidence of a grading bias in teachers. The perspective one should adopt depends on one's prior benchmark—are teachers assumed to be accurate (perspective 1), or is the algorithm assumed to be accurate, and teachers biased (perspective 2)?

[156] Recall that the purpose of the algorithm was to moderate the year-on-year increase and avoid high grade inflation. That moderation was *less* for private schools, and inflation was *more*, meant that even if private schools also saw some downgrading of predicted grades, pupils at private schools were more protected from downgrading. This may have been due to both the algorithm and the school—the algorithm may have biased private schools given it used school-level data, which is likely distorted by the economic advantages which are correlates with private school attendance, while private schools may have

had long been criticised for the outsized advantages they provided for students, as reflected in, say, the disproportionate representation of those privately educated in the highest paying jobs in the UK. There was a sense, then, of an algorithm containing, recreating, and thus perpetuating, unfair social dynamics—one might say an institutional discriminatory bias in favour of the wealthy.

The backlash prompted back-pedalling from the UK Government. Within days of the grades being released, the Department of Education announced that students would be given the higher of either their teacher predicted grades, or their algorithm predicted grades. As data from Ofqual (2024) show, this did indeed result in a grading bias—the percentage of students receiving the top A* grade doubled between 2019 and 2020; those receiving A and B grades increased by around 13% and 15%, respectively.[157] In 2021, where teacher's predictions were again used, overall grades reached an all-time high. This is against a historically flat trend for all grade boundaries for the decade prior to the pandemic.

The well-meaning behavioural scientist might be confused at this whole event. The data shows compelling evidence that teachers exhibit a grading bias, likely due to their inherent desire for their students to do well, and due to their tendency to recall the student's achievements rather than their failings. They might recall instances of previous students who did well, and who share some salient trait with the current student under consideration.[158] And so on. Given inflated grades have important knock-on effects—for instance, placing greater strain on more competitive higher education institutions[159]—a behavioural scientist may support the use of an algorithmic model or predictive AI tool to assign grades, at least in this special instance where students could not feasibly complete the exams.[160]

suffered even *larger* grading bias owing to the institutional dynamics of smaller classes and networked families.

[157] Note that a 'doubling' of those receiving the highest A* grade is only an increase, year-on-year, of around 7%.

[158] In terms of more 'fundamental' biases, one might label these the planning fallacy, the availability heuristic, and the representativeness heuristic, respectively. See Kahneman and Tversky (1982) and Tversky and Kahneman (1974).

[159] For instance, the University of Durham found itself having to accept more students than it had capacity for, resulting in the University paying some students to defer taking their place for a year (Weale and Adams, 2020).

[160] All of this raises an intriguing question about the future of examination and assessment. With the rise of generative AI, higher education institutions (and other education institutions) are very worried about the integrity of their assessments. Given the difficulties in outright banning the use of AI (beyond questions about the utility of doing so), and the unreliable nature of AI detection tools, the most intuitive recourse to this problem is to change how students are

Similarly, a behavioural scientist could analyse the narrative of the backlash and offer arguments to defend the use of an algorithm. For instance, emphasising the success of superstars reflects our difficulties with small probabilities, and might have led the British public to believe that those who beat the odds of their circumstances are greater than the odds would statistically suggest. There may be an element of loss aversion or the endowment effect, too, or even the conjunction fallacy.[161] A behavioural scientist could thus argue that the backlash to the algorithm was not driven by concern around algorithmic *errors*, but by biased assessments of a broadly accurate statistical model. Even on the matter of positive discrimination for private schools, a behavioural scientist may wash their hands of it, proclaiming this to be a d-bias, as above. Thus, from a behavioural science perspective, the use of an algorithm here is legitimate; opposition is biased.

Yet opposition to the algorithm, whether it can be explained by cognitive biases or not, came largely from a place of the public's *social values*—the *feelings* of people about what is right, what is fair, and so on. While there might have been costs associated with grade inflation, so too would there have been costs associated with algorithmic determination, and while one might suggest that the public did not fully or rationally consider the implications of their

assessed. One possibility, as is often done on MBA programmes, is to assess students at the end of each class session, over the course of the whole programme, based on their individual contributions and teamworking during the session. Though, this is often inhibitive in large cohorts. Furthermore, it may disadvantage those whose skills are not in public speaking and intense teamworking. As such a whole range of skills might need to form part of any future assessment—but this simply compounds the problem of their often being too many students to assess, and too many data points to consider.

Thus, one idea—though by no means an endorsement—might be to use AI to analyse student performance (measured in many ways, to capture many skills), and generate a recommended grade for an instructor, who may then use their discretion to adjust this grade. The simplest way to implement such a system would likely be an after-class question and answer session between an AI chatbot and a student, where the quality of answers as well as their accuracy is considered.

Though, as discussed below, such a recommendation obscures perhaps a more important point—if there are too many students for teachers to provide such one-on-one interactions, perhaps the solution is not a lack of technology, but a lack of investment in education.

[161] See Dhami and Sunstein (2023) for a discussion of poor human judgement when dealing with small probabilities. Loss aversion may explain the backlash as people seemed to respond more to the downgrading of grades, rather than to the upgrading. Endowment may be involved as parents were likely to consider their child more deserving and capable of higher educational outcomes by virtue of being *their* child. The conjunction fallacy may have been involved as people associate 'children' and 'teachers' with goodness, innocence, honesty, and so on, while 'government' and 'algorithms' may have been associated with error, inefficiency, and dishonesty. These are all my speculations, but speculation is the exercise here.

opposition to the algorithm, the critical question at this juncture is—*so what?*[162]

What this example demonstrates is the limits of behavioural science as a means of advocating for algorithms or AI in society—the *value-bias* problem. While one might use behavioural science to rationalise the activities both of public professionals and the wider public of which they are a component, this does not, in itself, change anything about the legitimacy of the underlying values expressed therein.[163] The use of algorithms is inevitably *political*, and there is a danger that in appealing to the 'inherent', 'systematic' biases of individuals to justify their introduction, the space for a political discourse is eroded.[164] As Bryne, Theakston and Randall (2020, para. 5-6) described the grading algorithm scandal within a wider political context, "Not only did [the UK Government] go down the algorithm route, but there was virtually no debate as to whether this was even wise—or just—beforehand. This speaks to the fact that the ability of algorithms to produce impartial, objective knowledge is now taken for granted in British political life. Algorithms, though, are inherently and inescapably political."

Selective Adherence and Machine Laundering

An interesting perspective emerges when one considers whether algorithms and AI technologies actually debias people. Snow

[162] Skidelsky and Skidelsky (2012, p. 155) perhaps pose a similar question when they write that, "Today, health is the one good on which liberal states feel entitled to take a positive stance, for, unlike the goods of the soul, it carries the authority of science. But is there really a distinction here? Science can tell us whether drug *x* treats condition *y*, but not that condition *y* itself constitutes 'ill-health'. This latter presupposes a pre-scientific, common-sense understanding of what it is for human beings to flourish."

Likewise, one might look to Feyerabend (1978, p. 137) who notes that Eastern traditional medicine, though 'non-scientific' by Western standards, often observes a normative injunction to not cause pain or to violate the body (e.g., through surgeries or invasive tests), which are frequently violated in Western medicine. For someone who holds these normative values dear, it may matter little that one is more 'scientific' or 'rational' than the other. As Feyerabend (1978, p. 10, original emphasis) notes, "Problems are solved not by specialists (though their advice will not be disregarded), but by the people concerned, in accordance with the ideas that *they* value and by the procedures that *they* regard as the most appropriate."

[163] Simon (1997a, p. 279): "When it is recognized that actual decisions must take place in some such institutional setting, it can be seen that the "correctness" of any particular decision may be judged from two different standpoints. In the broader sense it is "correct" if it is consistent with the general social value scale—if its consequences are socially desirable. In the narrower sense, it is "correct" if it is consistent with the frame of reference that has been organizationally assigned to the decider."

[164] There is some interesting evidence which suggests that public acceptance of algorithms in the public sector depends on citizens' trust of officials, and the contexts in which algorithms are used. See Ingrams *et al.* (2021), Kozyreva *et al.* (2021), Longoni *et al.* (2023), and Wenzelburger *et al.* (2024).

(2021) interviews several policymakers about how algorithms are used within policymaking. From these interviews, they document what they call *artificing*, or the use of algorithms and AI alongside one's own judgement. Artificing is primarily what behavioural scientists encourage when they encourage the use of algorithms in decision-making (Sunstein, 2024, 2023, 2022b, 2019). Artificing may be understood as a sliding scale bounded by two extremes. On one end, there is automation bias, or the tendency to use algorithms even when algorithms lead to errors.[165] On the other end, there is algorithm aversion, or the tendency to ignore algorithms even when algorithms lead to superior outcomes.[166] These are both interesting behaviours, with automation bias being the subject of some discussion in Chapter 5. However, the immediate discussion will focus on those behaviours which fall around the middle of this artificing spectrum.[167]

[165] Automation bias is often the subject of concern amongst the public (Russell, 2019), but the empirical evidence is mixed. Early studies of automated systems showed that people can come to rely too much on these systems, leading to errors which they would have otherwise not made (Mosier *et al.*, 1998; Skitka *et al.*, 2000, 1999). Some more recent evidence focusing on AI does not support automation bias (Alon-Barkat and Busuioc (2023).

Automation bias may also be likened to what has been called the 'Google effect,' or the tendency to forget information and other cognitive functions when these functions are given over an automated system. The notion of the Google effect is that information becomes less cognitively malleable when it is readily available (Sparrow *et al.*, 2011). Similar arguments were proposed about navigation skills from the introduction of satellite navigation, with popular headlines of people unthinkingly driving into rivers and lakes at the behest of their navigation system being common. Such a debate was also raised in antiquity, with Plato describing the meeting of the mythical King Thamus of Egypt and the inventor of writing, Theuth (or Thoth). Thamus, upon receiving Theuth and his invention, responds, "this discovery of yours will create forgetfulness in the learners' souls, because they will not use their memories; they will trust to the external written characters and not remember of themselves."

[166] Algorithm aversion is generally not dealt with as the focus here is on how people *use* algorithms, not on why they do not. Of course, rejecting an algorithm when an algorithm could produce a substantial benefit is a meaningful behaviour, and certainly should impact the arguments of behavioural scientists encouraging the use of algorithms and AI.

To briefly address algorithm aversion, studies across a variety of domains suggest people do sometimes avoid algorithms, even when it is in their interests not to (Jussupow *et al.*, 2020; Mahmud *et al.*, 2022), though much of the empirical evidence is lab-based, and may not reflect how professional decision-makers actually approach algorithms. Stevenson (2018) documents that judges given algorithmic decision-aids quite quickly stop using them, while Sunstein (2023) argues that algorithms are commonly used as second nature within hospitals (though the examples Sunstein gives may not be what most people think of when one thinks of an 'algorithm').

Some, such as Logg *et al.*, (2019) argue that evidence for algorithm aversion is exaggerated, and that evidence for algorithmic *appreciation* can also be shown. Others, such as Zhang and Gosline (2024), argue that it is not so much that people are *averse* to algorithms; rather, people merely *favour* humans.

[167] Meijer *et al.* (2021, p. 837) offer the terms "algorithmic cage" and "algorithmic colleague" which one might also use in this discussion. Though, these terms are

One result, supported by a growing body of empirical evidence, is that while different professional decision-makers *do* artifice, their behaviour is best described in terms of *selective adherence*—the tendency to selectively follow algorithms based on some irrelevant criteria—and specifically *confirmation bias*—the tendency to seek out, overweigh, and act upon evidence and information which conforms to what one *already* believes. Confirmation bias may be understood as a kind of selective adherence.

Alon-Barkat and Busuioc (2023) investigate the use of an AI algorithm in hiring decisions. Participants were tasked with evaluating candidates for a potential role, with an AI algorithm offering suggestions and recommendations throughout the evaluation process. The researchers find evidence that participants do use their discretion to overrule algorithmic recommendations. That people overrule algorithms should not be a concern. Indeed, in instances of ambiguity or uncertainty, discretionary judgement may be critical to ensure equitable outcomes (Sunstein, 2023). Thus, discretionary use of algorithms is not necessarily demonstrative of biased judgement and may reflect genuine insight which the algorithm lacks.

However, Alon-Barkat and Busuioc report that the criterion participants use to select when to adhere to the algorithm is when the algorithm's recommendation aligns with common racial stereotypes. Such a criterion is unlikely to lead to equitable outcomes and may in fact produce discriminatory outcomes; it is unlikely to be relevant to a hiring decision. Interestingly, Alon-Barkat and Busuioc do not just examine adherence to AI algorithms, but also to HR experts. Participants demonstrate the same selective adherence in both instances, suggesting that rejection of the algorithm is not especially driven by algorithm aversion.

Selten *et al.* (2023) investigate how police officers use algorithms when making resource allocation decisions. The researchers asked participants to decide from which of two locations to deploy police resources to resolve a robbery. Within the scenario, both locations would have advantages and disadvantages. For instance, deploying near the site of the robbery would increase the chance of a quick arrest without an ensuing, dangerous car chase. It would, though, risk letting the robbers escape if they evaded these early attempts. Deploying further away

less applicable to the idea of a spectrum of behaviour. The algorithmic cage is comparable to automation bias, and in some ways, may be a preferable term (see Chapter 5). The algorithmic colleague is comparable to artificing, though it is not specific as to the degree of collaboration it captures.

would reduce this risk, at the expense of heightened risk to the public.

Selten *et al.* (2023) find that police officers do not show any particular aversion or commitment to the recommendation algorithm. This is to say, they find no compelling evidence of algorithm aversion nor automation bias. However, when police officers *do* follow the algorithm's recommendation, it is overwhelmingly when it agrees with what officers *already* want to do. Conversely, officers overrule the algorithm when their prior views as to what should be done contradict it. As above, that officers demonstrate discretion in following the algorithm is not a matter of concern, as there may be genuine reasons for thinking the algorithm has made a mistake or missed some relevant detail. But that the criterion through which officers adhere to the algorithm is alignment with prior beliefs suggests that officer discretion is not used in a way which promotes equitable outcomes. Rather, it is used in a way which might undermine the benefits of the algorithm.

In another study of selective adherence, Narayanan *et al.* (2023) asked participants to evaluate case studies of patients seeking a new kidney and tasked them with making a recommendation to either approve a transplant, or to deny it—an ethically difficult, and socially important, decision. Participants were equipped with an AI algorithm to support their decision-making. Some participants received an algorithm that had been given a set of ethical preferences which differed from the participants; others received one which was congruent with their ethical preferences. In principle, the decision as to whether someone should receive a kidney should be based on objective evaluation of the relevant benefits *versus* the relevant costs of doing so (assuming a scarcity of organs), a function that an AI algorithm may be well-equipped to perform and may be valuable for a novice to have when facing such ethical quandaries.

Indeed, one might expect that because participants are amateurs, adherence to the AI algorithm would be high for all, indicative of automation bias. However, Narayanan and colleagues find that participants are more likely to follow the recommendations of the AI algorithm which aligns with their ethical views. The researchers find that such adherence is not because participants *feel* the AI system aligns with their views—it is not that the congruent system makes 'better' arguments. Rather, because the congruent system makes arguments that align with participants' views, advice is framed in a language and structure which the participant *already* agrees with, prompting adherence.

Nazaretsky *et al.* (2021) investigate selective adherence and confirmation bias in AI usage amongst schoolteachers. Across a series of interviews with teachers who had utilised AI technologies within their classrooms, the researchers find that teachers advocate for using AI, and adhere to AI recommendations, when the technology aligns with their prior beliefs, intuitions, and past experiences. When there was misalignment, teachers would generally reject AI technologies. In particular, there is a tendency amongst teachers to praise AI technologies in general but to be sceptical of their applicability in specific instances—say, in the context of *their* classroom and *their* students. This supports a confirmation bias interpretation insofar as teachers are more likely to hold prior beliefs about their specificities (the domains that they are experts in), but not about non-specific areas.[168]

Finally, Bashkirova and Krpan (2024) investigate the use of AI triage recommendations by psychologists working in the domain of mental healthcare. They find that the psychologists tend to reject AI recommendations only when recommendations do not align with their initial diagnoses and professional intuitions. When recommendations *do* align, psychologists tend to both follow the recommendation and trust the recommendation more. Unlike previous studies, though, Bashkirova and Krpan examine the impact of perceived expertise on adherence behaviour. They find that those psychologists who perceive themselves to have greater experience and expertise are significantly less likely to follow or trust the AI recommendation. One explanation for this result is that experts have more specific or detailed diagnoses, making it less likely for the AI recommendation to sufficiently align with the expert. Bashkirova and Krpan note that perceived expertise may frustrate the introduction of AI algorithms in professional decision-making domains.[169]

These studies are fascinating, and raise an important question for those who advocate wider use of AI algorithms in decision-making: *is a debiasing algorithm worthwhile if the algorithm itself will be used*

[168] Interestingly, in addition to this confirmation bias result, Nazaretsky *et al.* (2021) also report that teachers demanded a high degree of control over AI technologies—possibly to maintain their discretionary power in relation to their confirmation bias. Furthermore, they hold AI technologies to an extremely high standard—either predictions are perfect, or the technology is not considered reliable.

[169] Note an interesting, but unexplored, link between the role of expertise in artificing and the role of knowledge in the persuasion-knowledge (PK) model discussed in Chapter 3. This link reveals the benefit of conceptualising artificing as a spectrum rather than an absolute. Under the PK model, one's susceptibility to persuasion is influenced by one's topic knowledge. Where one has high topic knowledge, the PK model predicts high metacognition and high resistance to persuasion (Moon, 2010). The above studies on selective adherence suggest that one's expertise influences adherence to AI recommendations, with high expertise corresponding to high rejection of AI (low adherence).

in a biased way?[170] A related question, to which I now turn, is a question of motivation: *is selective adherence the consequence of cognitive bias which decision-makers would readily correct if informed, or might it be motivated by personal interests and institutional constraints*? The above studies do not offer a clear consensus on this question. Some (e.g., Alon-Barkat and Busuioc, 2023; Narayanan *et al.* 2023) suggest that selective adherence arises through implicit biases which the decision-maker may not be aware of and may be willing to change if informed.[171] Others (e.g., Bashkirova and Krpan, 2024; Nazaretsky *et al.*, 2021; Selten *et al.*, 2023) suggest that there may be professional prestige and expertise on the line when collaborating with an AI algorithm. One might speculate that in these instances, people would *still* selectively adhere but propose *post hoc* rationalisations for their decisions when challenged, because their behaviour is motivated by personal interests.[172]

A potentially useful framework to explore motivations for using, and overruling, algorithms can be found in Mills and Sætra's (2025) work. They ask whether AI systems can help policymakers make decisions which are more inclusive and representative. Focusing on sustainability and climate change, Mills and Sætra note that these issues rarely have obvious answers, and often involve both value judgements about what should be prioritised, as well as value judgements about value *systems*, as decisions often involve communities with different cultural and philosophical perspectives on value.[173] Decision-makers thus face an enormous information synthesis challenge (how does one consider such a huge range of perspectives?) and the challenge of competing interests (how does one balance a multitude of competing interests and claims?).

Poor representation and a lack of inclusivity arises when policymakers fail to overcome these challenges. Mills and Sætra propose a simple framework for thinking about the origins of such failure, which they call *categories of omission*. Firstly, decision-makers may be overwhelmed by too much information, succumbing to a suite of attentional biases and simplifying heuristics which cause

[170] This is posed as an open question. A reader is invited to consider it themselves and reach their own conclusion. In my opinion, that algorithms may be used in a biased manner is not a killing blow for algorithm advocates, but rather, a vital call for nuance, and a rallying cry against treating algorithms as a panacea for decision-making difficulties. While I am sure this is not a perspective genuinely held by many, in the realms of advocacy nuance can be lost, and minor benefits elevated through the unfair depreciation of important concerns.

[171] Though, no study to my knowledge has tested whether selective adherence continues after a participant has been informed that they are selectively adhering.

[172] Again, no study to my knowledge has tested this, though Bashkirova and Krpan (2024) offer similar musings, suggesting that the expertise of decision-makers makes it harder to convince them of their confirmation bias.

[173] One could readily draw parallels between climate debates and domains including criminal justice, medicine, education, and more. Indeed, much of public life is assembled around these areas of ambiguity and conflict.

some perspectives to be overlooked. Though, in principle, decision-makers would not choose to ignore a perspective if biases could be overcome. This category of omission is called *forgotten, but not opposed*.

Secondly, decision-makers may consciously choose to exclude a perspective because of personal self-interest and institutional and political constraints. Decision-makers still can (and will) suffer from biases; but even if these biases were overcome, perspectives would still be overlooked. This category of omission is called *opposed, whether forgotten or not*. Through this framework of categories, Mills and Sætra evaluate whether AI can promote greater representation, or not. The question of representation and AI is of most immediate relevance to a discussion of selective adherence because the decision to ignore (or not) an AI recommendation is comparable to the decision to ignore (or not) a novel perspective.

Mills and Sætra argue that AI technologies may be beneficial for overcoming omissions that are *forgotten, but not opposed*. Predictive AI could be used to synthesise information that the decision-maker may overlook, leading to a recommendation which is more accurate than the biased prediction the decision-maker will make (e.g., Sunstein, 2023). Predictive AI may even *automate* some aspects of a decision (Sunstein, 2024), leaving decision-makers to focus on more important aspects of representative decision-making—for instance, interpersonal interaction with different groups. Generative AI might be used to simulate underrepresented groups through silicon sampling, as discussed in Chapter 3. It may also be used to find novel and creative solutions to dilemmas which decision-makers cannot presently overcome, due to the many competing interests which must be satisfied (Bouschery *et al.*, 2023). Generative AI may even be able to foster more inclusive decisions by enabling many different people to submit perspectives and arguments, before summarising these submissions for decision-makers, decreasing the quantity of information while increasing the quality (Špecián, 2023). In sum, because AI technologies supplement the cognitive power of individuals in various ways, these technologies may ameliorate the factors which cause a person to overlook an alternative perspective, reducing omission and promoting inclusivity and representativeness. In some ways, this is a restatement of arguments for using algorithms to support professional decision-makers who exhibit biases (e.g., Sunstein, 2023).

However, when one considers the *opposed, whether forgotten or not* category of omission, Mills and Sætra are much less optimistic about the benefits of AI technologies. They note that there are many reasons why a legitimate perspective might be excluded which cannot readily be explained by the cognitive biases of a decision-maker.[174] The personal interests of powerful decision-makers, such as elected officials, can have a substantial impact on which ideas are considered, and which are not (Kingdon, 2003). Furthermore, the framing of the problem for which a solution must be found influences what perspectives are considered legitimate (Blyth, 2013; Hall, 1993), with decision-makers often consciously choosing to frame problems in terms of their personally favoured solutions (Feyerabend, 1978; Kingdon, 2003). Institutional factors can be substantially important, too, with budget constraints and past decisions impacting what a decision-maker can *actually* do (Kingdon, 2003; Simon, 1997a). An expensive proposal is likely to be omitted *by default* if a decision-maker only wants cheap solutions, while a cheap solution may be given substantial attention, even when it is demonstrably inadequate, *because it is cheap*.[175] Similarly, sunk costs in fossil fuel infrastructure are often cited as reasons for not divesting from these energy sources (Pettifor, 2019; Stiglitz, 2024)—despite overwhelming evidence for anthropogenic climate change.

These factors which lead decision-makers to ignore perspectives are not *cognitive* biases. Instead, they are political and institutional factors.[176] This means that even if one could overcome the biases of a decision-maker—which they will have—a perspective would *still* be ignored. Crucially, Mills and Sætra argue that AI technologies do little to overcome these political and institutional factors, and thus, when perspectives are *opposed, whether forgotten or not*, algorithms are unlikely to be very useful. In the case of selective adherence and confirmation bias; because overcoming these biases requires overcoming one's own ego and desire for prestige, one might be sceptical of whether algorithms will actually be useful.

Perhaps more concerning, though, is why one might choose to *use* an algorithm *despite* the political and institutional factors which constrain decision-making. Mills and Sætra argue that one might adopt AI technologies, despite their inadequacies in overcoming institutional and political barriers to meaningful

[174] The question of 'legitimacy' is obviously a relevant factor in this discussion, too. Though, in this instance, when I refer to a 'legitimate perspective' I simply mean any perspective an average person considers sensible. For instance, the suggestion that climate reparations should be calculated using a random number generator would not be a sensible, thus legitimate, perspective.

[175] For a thematically suitable study of how these factors influence policy adoption, see Mills and Whittle's (2025) study of the political-economic factors which influenced the rise of nudging.

[176] One might call these biases, or even d-biases. Though, this might inappropriately extend the definition of a d-bias. Furthermore, it may allow behavioural scientists to wash their hands of the implications which are articulated in this section and this chapter.

solutions, because one wishes to *appear* to be solving a problem. In the case of representation and inclusivity, an AI algorithm may allow decision-makers to argue that they are trying to foster more inclusivity (say through simulating diverse groups), without *actually* having to tackle the causes of exclusion (say, by inviting underrepresented groups into the room where decisions are made).

One might raise similar concerns about the use of algorithms in some areas of public and private life—for instance, when a professional decision-maker uses an algorithm only when it conforms to their prior beliefs and appeals to the 'objectivity' of the algorithm to resist opposition to their decision.[177] Mills and Sætra describe such uses of AI technologies as *technosolutionism*, following Morozov's (2013) use of the term to describe the use of technology to give the *appearance* of solving a problem, without actually solving it (Sætra and Selinger, 2023). When this is done for one's own self-interests, or for political or institutional reasons, one might call it *machine laundering*—the use of an algorithm to justify or hide what one already wanted to do.[178]

There are a growing number of examples which might be understood as machine laundering. In the Netherlands, the use of an algorithm to determine who should receive a childcare benefit resulted in tens of thousands of citizens being falsely accused of fraud. As Geiger (2021) argues in their discussion of what is now known as the *toeslagenaffaire*, political desires to reduce the benefits bill (austerity) combined with an anti-immigrant rhetoric, which often accused immigrants of falsely claiming state benefits, to create a political and institutional environment which *desired* the penalising of claimants. The algorithm laundered this desire, transforming this political objective into a technocratic, *objective*, determination of an algorithmic system—one which apparently did not suffer from the *subjective* biases of people who might otherwise be tricked by canny fraudsters.

[177] Stiglitz (2024, p. 242), for instance, describes what he calls a "façade of inclusiveness." Writing on international trade deals, he notes that, "Developing countries have demanded to take part in crucial global agreements because they have realised that if you don't have a seat at the table, you may be on the menu. But having a seat at the table isn't enough. Too often, their microphone has been effectively turned off, and no one is listening."

[178] The term machine laundering (again, the reference eludes me) could also be used in a discussion of copyright and generative AI. Insofar as training a generative AI system to output the 'average' of many copyrighted works obscures specific copyrights and disenfranchises copyright holders, generative AI could be described as a kind of laundering system. This perspective puts a different emphasis on the idea to that discussed here, but I do not think an inconsistent one. Insofar as generative AI may facilitate the violation or cheapening of copyright in a manner that someone *already* wanted to do; the system can be blamed for this outcome, rather than the individual or organisation. I generally will not deal too much with the question of copyright, but it is reasonable to involve it in a broader discussion of machine laundering.

A remarkably similar scandal, the *robodebt* scandal, unfolded in Australia in 2016. Here, an automated system erroneously accused thousands of benefits claimants of being paid too much by the government and demanded these 'debts' be repaid. As Pearson (2020) reports, the political landscape of Australia at the time (like the Netherlands) was one where politicians *already* wanted to cut the benefits bill, *already* believed in widespread fraud, and *already* treated benefits recipients as political scapegoats.[179] The algorithm simply laundered these political objectives, transforming them from political objectives and into 'objective' facts about benefits claimants—until the errors of the system were discovered.[180]

These examples are infamous because they involve erroneous algorithms, causing substantial harm to the lives of thousands of innocent people. One might object that these examples should not be applicable to a discussion of AI and behavioural science. The behavioural science argument is that algorithms should be used because, when designed correctly, algorithms and AI systems can be *more accurate* than people (Kleinberg *et al.*, 2018; Sunstein, 2023; 2022b). Yet, these examples are not given to highlight how algorithms make mistakes—though they do—but rather, to demonstrate how algorithms come to be *used* for a myriad of reasons, and how these uses intersect with behavioural science.

In the previous section, it was argued that the justification for using an algorithm to tackle current offence bias must be made with caution, as the bias label may justify an intervention which runs counter to our social values and ideals. But one might also adopt a machine laundering perspective, both to current offence bias and to mugshot bias. That people reoffend, and that their history is predictive of future offences, is demonstrative of a flawed criminal justice system which frequently fails to rehabilitate offenders. That judges place great weight on a defendant's mugshot is demonstrative of the implicit biases of judges and the fragile nature of the United States' diverse society.

[179] A tangential example may be the Horizon Post Office scandal in the UK. The UK's postal service, the Post Office, introduced a new IT accounting system, called Horizon. Horizon did not work, and produced accounting errors which appeared to show subpostmasters—those who ran Post Office branches—were stealing money. Thousands of subpostmasters denied being thieves, but the Post Office's senior managers did not accept there had been an error with Horizon and pushed forth with prosecutions of subpostmasters. A subsequent enquiry revealed that senior managers *already* believed subpostmasters were stealing from them. Their faith in the Horizon system, rather than in their employees, was in part driven by how the system confirmed manager's prior beliefs.

[180] An interested reader might glance at either (or both) Feyerabend (1978) or Illich (1973). Both have written extensively about how claims to objectivity are used for political ends, often for prestige and control. Also see Gramsci (2011).

If one wished to resolve these issues, substantial changes would have to be made to the criminal justice system; substantial investment would have to be given to American communities; and uncomfortable reconciliation would need to be undertaken to change the social mobility and cohesion between communities in the country. Such a programme would be tremendously politically ambitious—one might argue it is not a realistic programme. But one might also contend that it is not a programme which some in American society will desire. For instance, investment in criminal justice reform might require higher taxes or diverted spending in areas such as the military, which would likely see opposition from wealthier citizens and military contractors—powerful constituents in Washington! Labelling these flaws in systemic criminal justice as *individual* biases reframes the overall problem (Kingdon, 2003) into one which an algorithm or AI system appears to be a viable solution. This may be done because it allows policymakers to avoid alternative solutions which they *already* do not want to pursue.[181]

One can make similar machine laundering arguments about the other instances already discussed, involving doctors and teachers. For instance, a grading bias in UK schools may actually be demonstrative of a lack of teaching resources (e.g., time) to accurately assess a student's prospects.[182] Yet, given long-term funding constraints on schools, fiscal uncertainty from the pandemic and fiscal austerity in the years preceding it, providing

[181] Curchin (2017) and Mills (2024) have argued that behavioural biases may, in some instances, be understood as evidence of deeper social ills, and that continuing to treat them as biases has the effect of detracting from solutions to these social ills. This, to an extent, is an argument also put forward by Chater and Loewenstein (2023). Within the behavioural science community, this argument—which has been called the 'crowding out' argument—has been met with some scepticism. Though, in my opinion, much of this scepticism draws on a naïve model of policymaking and agenda setting. Mills and Whittle (2025) elaborate on the politics of nudging and behavioural insights in an attempt to demonstrate more completely the 'crowding out' perspective.

One perspective those who oppose the 'crowding out' crowd could present is an appeal to *realpolitik*—that some ideas are simply unrealistic within the political climate, and thus advocates should prioritise actions which are achievable at any given moment. Thaler (2021) presents a speckle of this argument when acknowledging that often what behavioural scientists can do is determined by politicians and prior legislation, and that behavioural scientists could be more ambitious if given the freedom to be. Still, an explicit *realpolitik* argument has yet to be made within the literature (to my knowledge), in part (I suspect) because it requires some concession to the 'crowding out' crowd of the role of politics and political influence.

[182] A recent study of nudges to influence food choices finds that time is a significant moderator of the nudge's effectiveness (Lohmann *et al.*, 2024). This suggests that time constraints play a role in biased behaviour—it may thus justify interventions not to counteract biases, but to alleviate their *causes* (Mills, 2024). Another recent study (Kaur *et al.*, 2024) finds that financial pressure, too, has impacts on cognition, specifically lowering worker attention and thus productivity. One could quite readily transplant these findings to jobs which often face financial constraints and time pressures—like *doctors* and *teachers*.

these resources may have been politically undesirable, while individualising the problem (as a bias) and solving *that* problem with an algorithm became a solution.[183] In the process, what decision-makers wanted (or did *not* want) to do was laundered by an algorithm, transforming a desired objective into an *objective* assessment of reality.

Machine laundering, then—the use of an algorithm to achieve some pre-held objective—is aided by behavioural science insofar as attributing a problem to an individual bias rather than a systemic failing transforms the solution from something politically difficult into something politically easier—an algorithm. Note here that behavioural science is not invoked as an appeal to some particular finding (e.g., as a science) but politically as an appeal to something 'objective.' Behavioural science is used as a *social technology* to transform how people see the world, and thus *act* within it.

What this chapter has thus so far attempted to convey is that the behavioural science of algorithms cannot be separated from the politics of algorithms. However, the behavioural science of algorithms can be used (purposefully or accidentally) to *hide* the politics of algorithms. This should be of concern to behavioural scientists. Those who are well-meaning and sincere in their efforts to help people and create a better world may hinder themselves if they are ignorant of the politics which surrounds them. Else, they may find their ideas tarnished within the political miasma. Calls for behavioural science to 'be humble' (Hallsworth, 2023) cannot just advocate humbleness from the behavioural scientist, but *advocacy* and *opposition* when behavioural science finds itself misused, or the *cause* of misuse.[184] Though this is not to offer a damning account of behavioural science's relationship to algorithms. The discipline could also be a means of *revealing* the politics of algorithms—for instance, by demonstrating the selective adherence and confirmation bias which influences how algorithms are used.

[183] As Weizenbaum once stated in an interview in 1985 (ben-Aaron, 1985, para. 22-24): "Why is there so much poverty in our world, in the United States, especially in the large cities? Why is it that classes are so large? Why is it that fully half the science and math teachers in the United States are underqualified and are operating on emergency certificates? When you ask questions like that, you come upon some very important and very tragic facts about America… It is much nicer, it is much more comfortable, to have some device, say the computer, with which to flood the schools, and then to sit back and say, "You see, we are doing something about it, we are helping," than to confront the ugly social realities."

[184] Efforts such as those of UK Behavioural Scientists (2020) to push back on the use of 'behavioural science' by the UK Government at the beginning of the COVID-19 pandemic are exemplar of what is meant by advocacy and opposition.

Your Colleague, the Computer

At the least, this chapter so far should suggest that even if algorithms can be used to reduce biases in decision-making, one might *still* be biased when using an algorithm. This also extends to the *initial* decision to *use an algorithm* or to *adopt an AI system*. This section explores this phenomenon through the link between behavioural science and technology adoption within organisations. Doing so provides additional, helpful perspectives to the arguments already developed in this chapter.

There is a famous saying within the world of management, only ever half remembered as it seems so obvious few give it substantial attention. It states that *one should measure whatever one manages*. It is often attributed to Peter Drucker, the famed management consultant, and is often invoked to reinforce the notion amongst managers that observation, surveillance, *information* and *data*, and so on, are essential to effective management. A manager who is not *measuring* cannot be *managing*.[185] Yet, Drucker never said *one should measure whatever one manages*, and the most likely source of this adage—Ridgway (1956)—did not argue that *effective measurement* meant *effective management*. Quite the opposite. Ridgway emphasises the dangers of quantification in organisations. By quantifying organisational phenomena, the messy, tricky qualitative components of, say, workplace morale, seemingly become a lot less messy. They are transformed into a much simpler problem of making a number go up, or down, which is often appealing to managers and, generally, to people.[186] Thus, drawing on Ridgway (and probably much more in the spirit of Drucker, too), one might rephrase the famous adage into something strictly more accurate: what gets measured *inevitably gets managed*.

Rather than being the proponent of quantified management, Drucker was a frequent critic, largely because he believed effective management was harmed by a recourse to numbers. While he did accept that manual activities could be managed through

[185] This is not really a new idea. Marx (2013, p. 1055, fn. 4) writes that, "The farmer cannot rely on his own labour, and if he does, I will maintain that he is a loser by it. His employment should be a general attention to the whole: his thresher must be watched, or he will soon lose wages in corn not threshed out… he must constantly go around his fences; he must see there is no neglect; which would be the case if he was confined to any one spot." Later, the Marxist management scholar Harry Braverman (1974) would emphasise the importance of Frederick Taylor's *scientific management* to 20th century capitalism.

[186] Peters *et al.* (2024) show that people are more likely to engage with social media content about climate change and climate science when numerical facts and data feature more centrally in the material. Though, too many numbers can be overwhelming, as Peters *et al.* (2007) show in a study of medical decision-making. Also see Peters and Markowitz (2024).

measurement,[187] he was sceptical that the same applied for knowledge work,[188] because such work involves much more exploration of immediate possibilities and challenges; it is a dance with uncertainty, rather than the march to the beat of a metronomic drum.[189] Thus, "because knowledge work cannot be measured the way manual work can, one cannot tell a knowledge worker in a few simply words whether he is doing the right job and how well he is doing it" (Drucker, 2006, p. 30).[190]

For the most important decisions, and the most transformational actions that an organisation might need to take, Drucker (2006, p. 143) entirely dispenses with the idea of measurement: "one does not start with facts. One starts with opinions." He continues: "To determine what is a fact requires first a decision on the criteria of relevance, especially on the appropriate measurement." This is hardly dissimilar—in fact, it is strikingly *similar*—to Simon's (1997a) distinction between *value* judgements and *factual* judgements.[191] For Drucker (2006, p. 145), to ignore the

[187] Drucker (2006, p. 2): "We have learning to measure efficiency and how to define quality in manual work during the last hundred years—to the point where we have been able to multiply the output of the individual worker tremendously."

[188] Drucker (2006, p. 3-4): "The imposing system of measurements and tests which we have developed for manual work—from industrial engineering to quality control—is not applicable to knowledge work… The knowledge worker cannot be supervised closely or in detail. He can only be helped. But he must direct himself, and he must direct himself toward performance and contribution, that is, toward effectiveness.

[189] Braverman (1974) makes the pointed argument that far too often the question of knowledge within organisations is simplified so as to diminish the contribution of workers. Braverman argues that while the scientist or the entrepreneur (the knowledge worker) might set into motion new ideas or new materials from which products may be developed; these workers fail to produce anything without the knowledge contributed the manual worker. Say, in figuring out *how* to implement an idea, or *how* to work with a new material. For Braverman (as for *Adam Smith*), the manager's central function in this flow of productive knowledge is to expropriate the manual worker's knowledge contribution (through surveillance and monitoring), and then control its *redistribution* back to the worker (through training, such as the rationalisation and simplification of tasks, and so on). For a more recent account, which links to Charles Babbage's political economy and his *difference engine* (as Braverman does), see Pasquinelli (2023, p. 84): "The epistemic imperialism of science institutions has obfuscated the role that labour, craftsmanship, experiments, and spontaneous forms of knowledge have played in technological change: it is still largely believed that only the application of science to industry can invent new technologies and prompt economic growth." To this end, one should consider again the *mechanical philosophy* perspective outlined in Chapter 0.

[190] Braverman (1974) disputes this point, arguing that all manual work was and is inevitably knowledge work, but knowledge work that has simply been atomised to the point of precise measurement. Thus, Braverman might contend that Drucker is merely distinguishing between work that has been effectively measured, and work which one has yet to discover an effective measurement for.

[191] Simon (1997a, p. 4): "Each decision involves the selection of a goal, and a behavior relevant to it; this goal may in turn mediate to a somewhat more distant goal; and so on, until a relatively final aim is reached. Insofar as decisions lead

role of opinions (or, in Simon's language, values) is to make a fatal mistake when important decisions must be taken.[192] He argues, quite compellingly, that "there would generally be no need for a decision" if the established measurement could respond to whatever challenge was now facing the organisation. That the organisation faces a challenge that a "simple adjustment" cannot respond to "indicates that the measurement is no longer relevant."[193]

The ills of the inevitable management of what is measured are important to the discussion of how algorithms should be used. This is because the decision to use of algorithm, or any technology, might be influenced more by a bias towards measurement, rather than through effective debate about how the algorithm *ought* to be used. A fascinating example comes from Simon (1981).

Simon reports a case involving the US State Department in the mid-twentieth century. This time period saw a confluence of factors which created a decision-making problem. Firstly, the Department received all their diplomatic cables via telegram, which were printed using teleprinters. These cables contained vital information which decision-makers needed to make effective decisions about what should be done, and what advice should be given to other parts of the US Government. Secondly, being at the height of the Cold War, diplomatic incidents were frequent and serious. Whenever there was a diplomatic incident, the Department would be bombarded with cables. Yet, the slow teleprinters could only print one telegram at a time. This frequently meant decision-makers faced hours-long delays in receiving vital information, harming effective decision-making.

Simon reports that the Department solved this problem *technologically*—they purchased more teleprinters. Because teleprinters could print in parallel, a doubling of the number of printers doubled the number of cables printed, halving the waiting time. Yet, Simon suggested this was unlikely to solve the

toward the selection of final goals, they will be called "value judgements"; so far as they involve the implementation of such goals they will be called "factual judgements." Unfortunately, problems do not come to the administrator carefully wrapped in bundles with the value elements and factual elements neatly sorted."

[192] Drucker (2006, p. 145): "Whenever one analyzes the way a truly effective, a truly right, decision has been reached, one finds that a great deal of work and thought went into finding the appropriate measurement."

[193] This is to say, where measurement is effective, it is the manager's job to monitor the data and to respond in an automatic fashion to deviances from some benchmark. For Drucker, decisions are not taken when data indicate output is down five percent, or employee turnover is up six. These deviances should automatically prompt action from the manager, in accordance with an organisational plan. If they do not—if the data simply prompt a questioning of what is to be done—then what is being measured must itself be questioned.

problem,[194] and that this solution came from a failure to truly examine how communication within the organisation worked. Simon (1981, p. 166) argues that the State Department focused on what they could readily measure—the number of cables printed—causing them to ignore what was actually causing the information backlog—decision-makers themselves: "A deeper analysis would have shown that the real bottleneck in the process was the time and attention of the human decision makers who had to use the incoming information." Thus, the *actual* solution would have responded to "a more sophisticated design problem: How can incoming messages during a crisis be filtered in such a way that important information will have priority and will come to the attention of the decision makers, while unimportant information will be shunted aside until the crisis is past?"

This would not have been an "easy problem" to solve. It would have, in Drucker's language, required a debate of opinion (what information matters?) rather than of facts (what does the information contain?). One might call such solutions *behavioural* solutions—solutions which involve the reorganisation of people. Yet, as above, because measurement and quantification simplify problems and convey a sense of control, while opinion and value judgements create the possibility for conflict (March and Simon, 1993), an apparent technological solution is often preferred to an actual behavioural one—in this instance, increasing the number of cables printed (Simon, 1997a).

This problem, which Mills and Spencer (2025) have begun to document, may be essential to understanding decisions about the deployment of generative AI. For instance, one study of publicly available computer code found that since the widespread adoption of coding AI 'co-pilots' in 2022, the *amount* of code written has increased significantly (Harding and Kloster, 2024). However, so too has the amount of code 'churn'—the editing and rewriting of code needed to make it work. This means that programmers are unlikely to be any more productive than they were before the co-

[194] Simon (1981, p. 167) does not report on whether the problem actually got worse, though he suggests that his proposed solution would have, "alleviate[d] the real problem instead of aggravating it," which implies that Simon was sceptical of more teleprinters being a viable long-term solution. This should make sense given Simon did not think this technological solution solved the *actual* problem. What is interesting is that such a solution is a classic case of Jevons' paradox. The failure of this solution likely came about because the extra capacity simply encouraged people to send *more* telegrams, clogging up the newly expanded printing highway. Jevons' paradox is most famously associated with traffic highways, where opening a new lane on a highway only temporarily reduces congestion; congestion returns because the new highway actually incentivises more driving, leading to more cars on the road. The solution here is rarely to build bigger highways, but to decrease demand for journeys and to increase supply of alternative transport options. See Duranton and Turner (2011).

pilot was introduced, because the *technological* solution means their time is now spent fixing broken code, while no *behavioural* solution has been introduced to resolve whatever might have been holding programmer productivity back to begin with. Though, what is readily measured—the *amount* of code—now appears to be increasing.

Other examples include a suite of summary and transcription tools now being introduced in existing software. One advertisement from Apple for their Apple Intelligence product shows a worker using AI to summarise a report they have not read when asked to take the team through it. In the context of the ad, *everyone* has read the report—what is really being asked of the (knowledge) worker is *what do they think*? Yet, AI allows the worker to *appear* to have read the report, while robbing the organisation of the original insights that that *specific* worker could provide, and in doing so, *destroying* rather than contributing value.

One might speculate as to why the worker did not read the report—were they too busy with other tasks, was their child sick and they could not find or afford child support, was the report simply not worth reading—and come up with *behavioural* solutions which allow this worker to show off their actual, unique and valuable insights. But doing so would require a confrontation within the organisation—are staff frequently overworked, should the organisation provide childcare provision, are senior managers too proud to recognise they are wasting people's time? Thus, a *technological* solution—AI summarisation—is offered, seized upon, and justified—in the mind of worker, and the organisation—by pointing to positive changes in whatever is measured.[195]

Another much touted application of generative AI technologies is in automating performance reviews (Jaffe *et al.*, 2024; Levy, 2024). Accepting for the sake of argument that performance reviews are actually worthwhile in the first instance (also see Graeber, 2015); the proposal suggests that workers could chat to a generative AI, which would then generate feedback and performance goals for the manager to rubberstamp at a later date.

[195] Mills and Spencer (2025) discuss this idea in the context of what they call *efficient inefficiency*. Efficient inefficiency arises when a technology is used to more efficiently undertake a task which is unnecessary no matter how efficiently it is performed. They argue that efficient inefficiency can create the *appearance* of productivity growth provided one believes the superfluous task is, in fact, necessary. In both the coding co-pilot example, and the AI summary example, there are apparent productivity boosts if one does not interrogate the nature of the task itself. Yet, despite these apparent boosts, Mills and Spencer argue this is still coming from an unproductive baseline. Simply eliminating the task would realise a greater benefit for the organisation. As such, technology is wasted when used in an efficiently inefficient way, while efficient inefficient itself is a *drag* on productivity.

The result is the manager may spend their time doing more productive tasks, while reviews are conducted faster.

The opportune question is, though, where does the value of a performance review come from? If it exists, it comes from the interpersonal interaction between worker and manager. It is a forum for the worker to voice their problems, the manager to speak candidly, and both to 'negotiate' a progressive path forward (March and Simon, 1993). That a report is written and signed at the end is wholly irrelevant to the value contribution of a performance review. But the number of reports is what is *measured* when the manager is themselves assessed. Thus, the problem is understood technologically, rather than behaviourally; questions are not asked of why reviews take so long, why a manager or worker might not have time for them, and so on. As such, the uncomfortable *behavioural* solutions are implicitly ignored, while the more comfortable technological *solution* is favoured because a) it avoids conflict; and b) it appeals to what is measured. That it also c) destroys the value of the exercise is less relevant.[196]

One might ask as to why this would be allowed to happen—what manager or organisation would sanction such uses? One answer comes from recognising that the organisation is not a single entity, but rather an assemblage of different people, groups, motivations and interests (Simon, 1997a). Individuals in organisations are themselves boundedly rational, and this character then echoes throughout the whole organisation (Cyert and March, 1991). To give but one hypothetical, an executive has to deal with conflicts arising from shareholders, trade unions, and middle managers. Being boundedly rational, the executive must prioritise and focus on the most serious conflicts. Those less serious conflicts—say, a middle manager's push to use AI summaries in meetings—may not receive an adequate level of scrutiny owing to a deficit of cognitive resources. The organisation may thus do things which are not optimal because executives must satisfice and place their priorities elsewhere (Cyert and March, 1992). Only the hypothetical, economically rational organisation avoids such issues, and no behavioural scientist should believe that such an organisation actually exists.

Adopters and advocates may also *lack* the knowledge to use technology well (Mills and Spencer, 2025). ChatGPT, for example, allows organisations to use AI technologies without employing those with programming skills. The net effect is an overall lowering of the skills needed to use technology, and thus, those found in the

[196] From this perspective, one might also bracket this as a kind of machine laundering. It is also essential to note that even if these are useful tasks to undertake, AI will only contribute to productivity if the released resource (e.g., time) is put to use doing something worthwhile.

organisation. This has short-run advantages for an organisation—for instance, lowering the cost of staff—but has long-run negatives in terms of efficiency and productivity. This is because organisations lack the valuable knowledge of how processes and technologies work, hindering the discovery of ways in which they *could work better* (Acemoglu and Johnson, 2023).[197]

One worthwhile example is the self-service kiosk. Self-service kiosks allow supermarkets to dismiss checkout operators. But these technologies do not replace the operator with a faster, more efficient machine. Rather, they merely *shift* who performs the task (Lambert, 2015). Now, instead of a skilled checkout operator, scanning is done by customers who lack the skills to make checking out a fast and efficient process. Simple problems, such as dealing with errors, cause long delays as the customer neither has the knowledge to fix the problem, nor the authority to implement the solution. And, even if the customer *did* find a more efficient use of the machines, or a way to speed operations up, they have no incentive beyond their own occasional convenience to share this insight. This use of technology, which one might call *technological disintermediation*, is increasingly prevalent in modern society. As above, ChatGPT technologically disintermediates the programmer, leaving the everyday user to stumble around and figure out what the technology can and cannot do, what it should and should not do, and so on. In such an environment, where shallow knowledge prevails, it should not come as a surprise that what is measured becomes what is managed, nor should it be surprising that comfortable *technological* solutions are preferred over uncomfortable, disruptive *behavioural* ones.[198]

[197] Economists call these skills *human capital*. Acemoglu and Johnson (2023) argue provocatively that technology only brings benefits when it enhances human capital. This can be achieved through *augmenting* technology. For instance, giving a drill to a skilled craftsperson enables that person to make more, high quality goods. When technology *automates* work, it destroys human capital. Now, the organisation could fire the skilled craftsperson and hire a lower-skilled machine operator. The organisation will see lower labour costs, while the machine might match the craftsperson's quality. But now the organisation lacks the knowledge of the craftsperson to innovate new products, new techniques, and new ideas about how the machine could be used. Thus, in the long run, the organisation will stagnate.

[198] One might consider how far this argument extends. Ours is a time punctuated by promises of technological innovation and terrible disappointment. Chapter 2 must give one pause as to the promise of targeted advertising and recommendation algorithms; to this list one might add the metaverse and blockchain. Some would suggest generative AI itself will soon join the list—hardly unjustified, if the misuses discussed in this chapter are realised. More controversially, some might even suggest the computer, given arguments that it has failed to substantially impact productivity (e.g., Acemoglu *et al.*, 2014).

The technological disintermediation argument extends to the organisation itself. The advent of Silicon Valley and a 'technology industry' is historically unique. Today, unlike earlier epochs, organisations do not draw on their own expertise and experiences to solve problems. Innovation has been

Contrary to what some readers might be inclined to believe, the purpose of this chapter has not been to argue that algorithms are *useless* in decision-making. Throughout this chapter, discussion has centred on the *misuses* of algorithms, accidental and intentional. But one cannot *misuse* something that is *useless*. To this end, a reader should not imagine the conclusion of this chapter is that algorithms have no place in decision-making.

Intelligence Is What We Make of It

People can be strange, and technology can make people stranger. After creating ELIZA, Weizenbaum developed a variant of the programme which would go on to become the most famous version of ELIZA, known as DOCTOR. This programme roleplayed as a therapist, responding to people to encourage a deeper exploration of one's inner psyche. Weizenbaum (1976, p. 6) notes how he, "was startled to see how quickly and how very deeply people conversing with DOCTOR became emotionally involved with the computer and how unequivocally they anthropomorphized it."[199]

How did ELIZA work? Essentially, ELIZA had a bank of text responses with blank spaces which could be filled with keywords from a person's message. Basic logic trees structured simple conversations that ELIZA might encourage, and often, the programme would simply rephrase a person's message as a question back to them. One might mention their sister. ELIZA might respond: "Sister?" One might then elaborate about their

outsourced to professional technologists, who invent the future. They are guided by their imaginations (hardly the worst guide) but lack the practical knowledge of what problems need to be solved precisely because they are separate from the organisations for which they are developing technology.

Such an arrangement works for universities, which specialise in the discovery of knowledge separate from an organisational setting or 'the coalface.' But universities do not exist to make a profit and focus on 'fundamental' or 'far' technologies; technologies which do not have immediate commercial applications. Universities socialise the risks of innovation and leave discoveries in the public domain for others to transform into 'near' or 'late' technologies. They thus complement the coalface, and do not suffer from being at a distance from it.

But today's technology industry both wants distance and commerciality. They want to sell the possibility of an organisation employing only those with the minimum-necessary skills to perform their task, and many organisations want to buy this story (Acemoglu and Johnson, 2023). But in doing so, organisations forgo the opportunity to spot innovations and adapt technologies, while technologists fail to acquire the deep organisational knowledge needed to develop actually useful technologies.

[199] Weizenbaum (1976, p. 7) continues: "I know of course that people form all sorts of emotional bonds to machines, for example, to musical instruments… What I had not realized is that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people." Note that Weizenbaum uses the word 'machine' where, per Chapter 1, *tool* is more accurate.

relationship with their sister as children. ELIZA might respond: "Tell me about your childhood?" One might mention it being a happy childhood. ELIZA might respond: "How was it happy?" And so on. In some ways, the simplicity of ELIZA makes it a beautiful programme, one that can and should be appreciated not as a chatbot or an artificial intelligence, but as a piece of socio-cultural engineering.

Despite the simplicity of how ELIZA worked, Weizenbaum (1976, p. 6) also observed, "Another widespread, and to me surprising, reaction to the ELIZA program was the spread of a belief that it demonstrated a general solution to the problem of computer understanding of natural language… [Using ELIZA] I had tried to say that no general solution to that problem was possible, i.e., that language is understood only in contextual frameworks, that even these can be shared by people to only a limited extent, and that consequently even people are not embodiments of any such general solution." He continues: "But these conclusions were often ignored. In any case, ELIZA was such a small and simple step. Its contribution was, if any at all, only to vividly underline what many others had long ago discovered, namely, the importance of context to language understanding… This reaction to ELIZA showed me more vividly than anything I had seen hitherto the enormously exaggerated attributions an even well-educated audience is capable of making, even strives to make, to a technology it does not understand."[200]

[200] For instance, Illich and Sanders (1988) in their history of writing and language, note that the first songs, poems, and stories demonstrate a composition which is inconsistent with modern writing, but reflects conversation and 'folk' development of narratives. Writing emerged *after* language, and while stated bluntly this is hardly surprising, the implications are often under-appreciated: there was a time when writing was alien to communication, and where language existed *only* as a social medium between people. Graeber (2015) would argue thus is still the case, only that we have deluded ourselves into thinking otherwise.

Amongst other things, Illich and Sanders note that with the advent of writing came the emergence of 'knowledge,' which was not knowledge of communities, or learned understandings of turns of phrase or the meaning of allegories; but knowledge embedded *in text*, and ultimately, *as text*. Citing Plato's dialogue between Theuth and Thamus, Illich and Sanders suggest such knowledge is ultimately less helpful, and possibly more harmful, than is presently appreciated. To Theuth, the mythical inventor of writing, the mythical King Thamus suggests that writing will, "give your disciples not truth, but only the semblance of truth; they will be hearers of many things and will have learned nothing; they will appear to be omniscient and will generally know nothing; they will be tiresome company, having the show of wisdom without the reality."

What one must appreciate in this argument is that writing can be removed from a social context, and in turn, become nothing more than scribbles on a page. Spoken language, while it can be *forgotten*, cannot be removed from a social context. Hence Thamus' argument that those who learn only through writing lack the social context to make the knowledge they gain have any meaning.

Today, we might consider ourselves beyond this point of ignorance. Ours is the information age, and every day more of us find our first identities, our first communities, our first loves and heartbreaks, through the computer and the internet (Turkle, 2013). But what is fascinating about ELIZA is not the trick it perhaps played on those who should have known better *then*; but how it colours the peculiarities of popular interpretations of AI and algorithms *now*.

Take, for instance, the notion of AI *hallucination*. Generative AI is said to hallucinate when it makes something up, or states something which is false by some objective measure of reality. But this idea obfuscates the reality of the situation. AI systems do not *know* anything; they do not think. In feats of astonishing engineering, sophisticated mathematics is used to subsample what is essentially an enormous database.[201] Words are generated sequentially based on a probability distribution. This is the process by which all diffusion-based outputs are generated. Thus, AI does not hallucinate; or, if it does, it *always* hallucinates. There is no output based on imagination or speculation, and there is nothing technical to distinguish a 'true' statement—which is meant to demonstrate intelligence—from a 'false' statement—which is meant to demonstrate hallucination.[202] The same is true of predictive AI, though here it is more subtle. Given enough variables and enough data, any number of statistical patterns can be found, and used to predict some outcome. Whether those patterns have any meaning in the prediction, and thus map onto

[201] That AI systems are simply databases is well articulated in Salvaggio's (2023) wonderful short film *Flowers Blooming Backward Into Noise*. Here, Salvaggio argues that AI generated images are merely infographic representations of the average of a database of images. Based on my understanding of diffusion models, this is technically accurate. It is also not a description many would recognise, as we call AI *artificial intelligence* rather than what is technically more accurate: a *diffusion engine*.

[202] As a fun exercise, ask ChatGPT or some other LLM to 'think' of a number between 1 and 10. Instruct it to not tell you what the number is until you clearly ask to be told. Finally, instruct it to answer 'yes' or 'no' to whatever question you ask it in relation to the number it has 'thought' of. One will probably receive a response such as, "Sure! Go ahead and ask me any yes or no question."

Say one asks if the number is divisible by three, and the response is "no." Then one asks if the number is a cube number, and the response is "yes." Finally, one asks if the number is 1, and the response is "no." The only mathematically correct answer is thus 8. But when asked to reveal what the number *actually* was, the AI system is as likely to say "8" as it is any other number. This is because when initially prompted to 'think' of a number, no thinking took place; nowhere in the system was a variable x assigned a number for the purposes of the game.

Instead, the system responded with a *probabilistically likely natural language response*: when asked to think of a number but not to tell you, most people will respond "Sure, I have thought of a number." But most people *will* actually think of a number, too! This exercise demonstrates that while people perform both a language function (i.e., in responding to you) they also perform a social function (i.e., thinking of a number, participating in the game). ChatGPT, and other AI systems, only simulate the language function.

some actual phenomenon in the world, cannot in itself be determined by the act of statistical pattern spotting, as noted in Chapter 2.

As Weizenbaum notes, in each instance, what gives these outputs meaning, and these systems 'intelligence' is *us*. People supply these systems with context and place these outputs in a social setting from which meaning can be constructed (Illich and Sanders, 1988). People determine whether an output is 'true' and thus intelligent, or 'false' and thus a hallucination. People decide whether a statistical pattern has meaning within the social and institutional setting it arises in, or not. As Pasquinelli (2023, p. 235) argues, intelligence is fundamentally a *social* phenomenon, something that emerges between people in the process of social interaction: "there is no inner logic to discover in intelligence, because intelligence is a social process by constitution."[203] AI enthusiasts, to an extent, understand the concept of emergence, recognising that complex system behaviour can arise through the interactions of individually simple agents. But this perspective diminishes the social component of emergence. This, in turn, hides the *anti-social* nature of diffusion systems—whenever an AI system 'hallucinates' it is attempting to automate the supply of context, to omit the human contribution to the construction of intelligence by filling in details which only come to be erroneous because people themselves have that context and can supply it.[204] In Marxist terms, the machine replaces *living* labour with *dead* labour, and insofar as AI systems are engineered to automate context and exhibit intelligence without sociality, the notion of *death* seems fitting.[205]

[203] As noted in previous chapters, this insight was likely not lost on Turkle (1988) or Minsky (1986), but in my opinion is made best by Weizenbaum. Pasquinelli's recent contribution is praiseworthy insofar as it attempts to recast some of these discussions against recent AI developments.

[204] Pasquinelli (2023, p. 234, original emphasis): "[P]erhaps the most important aspect of the [AI] classifiers has nothing to do with their *internal logic* but with the association of their output to an *external convention* that establishes the meaning of an image or other symbol in a given culture. Gestalt theory, cybernetics, and symbolic AI each intended to identify the *internal laws* of perceptions, but the key feature of a classifier such as the perceptron is to record *external rules*—this is, social conventions. Ultimately, an artificial neural network is an *extroverted machine*."

[205] For the interested reader, one can take this critique further. For instance, the notion of the market economy as a complex system or machine in which intelligence emerges through the individual interactions of people (e.g., Hayek, 1999, 1945) hides the fundamentally *anti-social* nature of the market—one that seeks to replace the sociality which governs the production of use-values with an automatic mechanism which governs the production of exchange-values, and which hides the commensurability of the two. Unsurprisingly, this idea has links to Marxist philosophy (Marx, 2013; Skidelsky and Skidelsky, 2012), particularly the notion of alienation—a detachment from the sociality of one's work and effort. See Braverman (1974) and Davies (2024).

There are also links to Illich and Sanders' (1988) critique of language and writing. As Skidelsky and Skidelsky (2012, p. 41) write of finance, "traders in

This all offers an important twist on the idea that algorithms should be used because decision-makers are biased. Not only do people give AI systems 'intelligence' through the meaning and social context we provide to system outputs; but in labelling people as biased, our own claims to authority and agency within the world become conditional upon the algorithm. Caution must precede the behavioural scientist.

futures, derivatives and other rarefied financial products need know nothing at all of the actual goods that lie at the end of their transactions. Living in a world of pure money, they lose feeling for the value of things." Recalling Chapter 0, it is interesting to again consider that modern AI systems essentially assume the relation between reality, R, and data, D, is given by some equation $R = \omega D$, where ω is a variable the AI system is designed to estimate, thus simulating R from D. In the same way, prices (P) and value (V) can be related mysteriously through the equation $V = \gamma P$, whereby the market is said to serve the same function as the AI in estimating γ. Knowing ω or γ supposes one does not need to know reality, or value. Only data, and prices.

Chapter [5]—The All Seeing I

> "[The] twenty-first century, in which the only knowledge that counts is prediction."
>
> —Jill Lepore, *If Then* (2021, p. 5)

Autonomous Choice Architecture

Previous chapters have generally dealt with use of AI systems, within behavioural science, as a tool. Increasingly, however, AI systems are being introduced as part of behavioural technology systems which function as machines, possessive of 'their' own motive power. To be clear, it is a common inaccuracy to say that 'AI does *x*'—rarely is it that AI performs the final function of whatever system is being examined. Just as the brain does not pick up items, AI systems do not act in the world, *per se*. In both instances, these decision centres are connected to sensory organs—cameras, eyes, (robotic) arms, and so on—which act in the world through signals from these centres.

Appreciating the role of these sensory organs is vital to fully grasp what it means to say that an AI system has motive power; that it is *autonomous* or a *machine*. A predictive AI system, for instance, has motive power insofar as neural networks function automatically once activated. But they may lack motive power once an output is generated.[206] The motive power of a model that predicts a car should turn left depends on whether that prediction automatically feeds into the steering of the car, or whether it simply pops up as a suggestion for a person who may always choose to go right, instead. In the former instance, the model is part of an autonomous system—a machine, specifically, a self-driving car. In the latter instance, the model is dependent upon the driver; it is a tool, specifically, a GPS.

Even more formally, still, sensory organs might be described as those parts of a system which interact with the external environment (Simon, 1981). 'Interaction' here means not just receiving a signal from a decision centre, such as an AI system, and responding to that signal. It can, and often does, mean supplying the decision centre with information in the first instance. A person's arm provides information about an object's temperature to the brain, which leads to a new signal that may or may not

[206] The same is true of generative AI. A generative AI, set into 'motion' (so to speak) automatically generates an output, but whether automaticity applies beyond the output depends on the motive power of the system as a whole.

prompt the arm to change what it is doing (Ashby, 1978). Similarly, a camera may feed images to an AI system which detects an intruder; the system signals that an alarm should be sounded; the camera signals images showing the intruder has retreated; the AI signals for the alarm to cease. While such a description might be pedantic, it demonstrates that motive power does not just mean action independent of human oversight but can also mean *reaction*.[207]

This chapter deals with a system described in these terms. When an AI system is equipped with sensory organs to design, deliver, and monitor choice architecture, such a system can be called an *autonomous choice architect*, and the behavioural technology itself, *autonomous choice architecture* (Mills and Sætra, 2024). This system works not merely through software and hardware, but also via insights from behavioural science. Furthermore, it does not work to move vehicles or assemble components, but to affect the behaviour of people within the world. It is from this perspective that behavioural science can be understood as a *social technology*, as noted in Chapter 2. This chapter focuses on autonomous choice architecture; how it changes behavioural science; and the social consequences contained therein.

Discussions of autonomous choice architecture can come close to debates in other fields, such as AI ethics and technology law literatures, to the point that this discussion could—on the one hand—become diffuse, or—on the other—be accused of being incomplete.[208] My approach is to address a common discussion within these literatures which is most applicable to autonomous choice architecture: *explainability*. While these adjacent literatures may offer further insights for a scholar of autonomous choice

[207] The above descriptions, and conceptual approach more generally, is based on the cybernetic notion of a *homeostat*, or a self-regulating system. While the systems to be discussed in this chapter are not to be taken as pure homeostats, that there is a component of feedback within these systems is essential. For the curious reader, this description and this chapter should be seen as further evidence of the link between behavioural science and cybernetics. See Ashby's (1978) *Design for a Brain*.

Much of what has been described above can also be found in various works by Simon (e.g., Simon, 1997a, 1981), who likens the above description of both a person and an AI system to that of an organisation. Indeed, the notion of a 'decision centre' used above draws directly on Simon's (1997a) use of the term in describing executive communication within organisations. A curious reader may thus also wish to draw connection to an earlier discussion about the conceptual similarities between individuals, computers, and organisations insofar as one is concerned with the question of 'intelligent' decision-making.

[208] This area has always suffered from difficult language and conceptualisation, which has often been the result of scholars seeking to differentiate those discussions concerning behavioural science and decision-making from those concerned with human-computer ethics more generally.

architecture, this discussion will then shift to topics more directly applicable to behavioural science.

Mills and Sætra (2024) argue that one could object to autonomous choice architecture on explainability grounds (also see Mills, 2022b). It is reasonable to suggest that choice architects should be able to explain their decisions around the architecting of choices, if ever asked to.[209] Such explanations do not have to be supported by each individual. Indeed, as the problem of heterogeneity demonstrates, often the rationale for nudging outcome A does not apply to person 1. As such, person 1 might not support the nudge. But support is not the same as explainability. Even where the explanation fails to garner support, there is an explanation which could be offered. For instance, that: *most people benefit from choosing A, and on utilitarian grounds, you have been nudged towards this outcome, though it may not be best for you, specifically.*

By contrast, an explanation may not always be forthcoming from an autonomous choice architect. Say two friends enter a restaurant where menus are personalised using autonomous choice architecture, following methods discussed in Chapter 2. The time is around 11:30am. Each friend uses their smartphone to access the menu, logging into their respective accounts with the restaurant. One of the friends is shown a breakfast menu, while the other is shown a lunch menu.

[209] The idea behind this assertion is one of a) democratic accountability; and b) dignity and respect. If one feels so disposed to nudge someone towards an outcome, one should also be capable of offering a rational argument for why that outcome should be chosen. That rational persuasion is not used, and choice architecture is, is a problem of context—perhaps rational persuasion is not a practically feasible strategy? But *if it were* practically feasible, one should be able to deploy it. If they cannot, the whole basis of advocating for that outcome can and should be questioned—indeed, if one nudges precisely *because* they lack a rational, persuasive argument, this could be considered a wholly illegitimate and manipulative strategy.

Relatedly, Sunstein (2014) explores instances of why people might, sometimes, choose not to choose when faced with a decision—an important but underexplored area of behavioural science. One reason—also acknowledged by Simon (1997a) and game theory scholars, such as Schelling (1980)—is that letting someone else decide allows one to exploit that person's superior knowledge and expertise. For instance, one might follow a default option because one recognises that whatever option is implicitly recommended to them is probably based on a superior assessment than they themselves could ever undertake. This is known as *information leakage* (McKenzie *et al.*, 2006; Sher and McKenzie, 2006; Sher *et al.*, 2022).

It is not always the case that advocates have superior knowledge. Furthermore, they may have malign motives. Thus, this is not a universal explanation for why people may defer to implicit recommendations, such as default options (and nudges in general). But that people do 'free ride' and use the advocacy of others as cognitive shortcuts when making their own decisions cannot be ignored (Sunstein, 2014). As such, in spheres such as public policy, advocacy should be made *at the least* for explainable reasons.

Confused, the pair usher over a waiter to ask *why they are being shown different menus?* The waiter makes various apologies and explains that everyone receives a unique menu. Their job is simply to deliver the food once prepared and the chef's job is simply to prepare the orders. The chef does not even handle inventory—an algorithm predicts what food the restaurant will need and automatically orders it. When the friends ask the next obvious question—*why did we each receive these specific options?*—the waiter explains to them that the options they are shown are those which have been algorithmically predicted to be most desirable to each of them individually.[210] Even the ordering of options on the screen; the colours of the menus; the descriptions of each dish; and so on; each feature of the menu has been automatically, algorithmically determined (Hauser *et al.*, 2010; Reinecke and Gajos, 2014).

Perplexed, the friends ask the final question: *how?* The waiter gives the final response before the awkward silence ensues—*I don't know, I just work here.*

In the context of a restaurant, perhaps all this is forgivable. But when one extrapolates this line—*I don't know, I just work here*—to other settings, be it to doctors in hospitals, teachers in schools, tellers in banks, or social workers or the police in governmental institutions, such a phrase takes on a more serious, if not *sinister*, character. To be sure, there are means of recourse, here. Certain jurisdictions, such as the European Union, have legislation implementing so-called 'right to explanation' procedures. These demand that any algorithmic decision be explainable to a person effected, upon request (Wachter *et al.*, 2017).[211] Yet, this is not a silver bullet. Edwards and Veale (2017) argue that the ambiguity around what it means to explain something makes this right essentially unenforceable. For instance, one might be told their gender contributed to a decision. If so, how much? If told that it contributed 31%, where did that percentage come from? And so on. Where does explainability end (Mittelstadt *et al.*, 2019)?

One might attempt to answer this question. For instance, an effective explanation is often one where the values of a judgement are laid bare before someone. Poor explanations often deal in technicalities and factual statements. Better explanations often deal in advocacy for what one believes, thinks important, and thinks *ought* to happen. To explain that '*you have been nudged because we believe*

[210] For the sake of argument, one should assume that the restaurant does have the best interests of the diners in mind, *as judged by themselves*. In reality, such a system would likely be optimised to maximise the spending of the diner. In *some* instances, the diner's preferences may also align with those of the restaurant (e.g., the diner wants the most expensive item). Though, this is likely a rare occurrence.

[211] There are also frameworks for 'auditing' algorithms, dissecting the various design decisions contained within automated processes to offer explanation of them (Raji *et al.*, 2020).

most people will benefit from this outcome' is a statement which deals more in values than in facts.[212] One may then come to their own conclusion as to whether these values are acceptable or tolerable; one does not need to seek more details before they can reach this conclusion (Simon, 1997a).

One of the challenges of explainability when considering automated systems is that these systems can rarely expose the values inherent in their design. Such systems seem disposed to dealing with facts. A model may be able to explain that gender contributed 31% to an outcome; but that model likely cannot explain *why* gender matters, or why it *ought* to matter. It can explain through only facts, rather than values.[213]

Following the Leader

Another reason why autonomous choice architecture may create difficulties in explanation arises from how automation influences the relationship between people and processes. Complicated machines are rarely *just built*. More commonly, simple tasks are automated through simple machines, which over time come to be 'stitched together' into a more complicated machine (or system of machines) where the once clear division of tasks is now lost in a continuous autonomous *process* (Braverman, 1974; Davies, 2024). Equally, automating activities formerly undertaken by people, the knowledge of those activities no longer needs to be reproduced among those people or those who follow them (Frey, 2019; Simon, 1997a). Over time, knowledge of activities becomes scarce, even as these activities *cum* processes proliferate.[214] Explanation is sacrificed for automated ends.

[212] This is not perfect. One could always ask: *why do you believe most people would benefit from this outcome?* The answer would often then be something about a cost-benefit analysis, or some other weighing of options—a technical, fact-laden answer. Though, further probing could then reveal the values contained within these more specific facts.

[213] This is a rephrasing of various arguments, made in Chapter 4, around machine laundering. That automated systems can *hide* values, often by transforming (as above) political objectives into objective facts, is not a bug, but a feature, in some instances, and reflects the challenges of explainability which arise when considering automated systems.

[214] The above paragraph adopts language which suggests people choose to give up knowledge to machines. Various scholars would attest that this is not the case (e.g., Braverman, 1974; Marx, 2013). The loss of language, for instance, is often the result of conquest. That written forms of language have, historically, been used by conquerors as means of control can be understood as an attempt to automate language through the technology of writing (Illich and Sanders, 1988).

The Luddites—those people famous for smashing up machines—can be understood in similar terms. Their objections were not to machines or the evolution of manufacturing *per se*, but that the machine created conditions whereby the Luddite's skills no longer needed to be reproduced, and thus the Luddite's way of life would no longer be supported (Braverman, 1974; Merchant, 2023). Modern anxieties about automation may often be compared to that of the

This objection, and other concerns about explainability, might be overcome if people retain oversight of autonomous choice architecture (Sunstein, 2023). The argument might go that so long as people have discretionary power to overrule AI systems, the explainability challenges which arise through the use of these systems disappears. This is because, regardless of whether discretion is exercised, the person with that discretionary oversight can explain why they did, or did not, allow the autonomous system to act autonomously.

Yet, it is worthwhile to consider that there may be many instances where people do *not* exercise discretion over autonomous systems, even when they technically could do so. Chapter 4 began to examine this question, to an extent, by considering the artificing behaviour of selective adherence. Yet, selective adherence considers the motivations for both adhering and *not* adhering to automated systems. It is immediately helpful to focus on the former in more detail, and ask *in what instances might people choose to not overrule an algorithm or automated system?* Broadly, there are three reasons.

The first may be a matter of convenience. Sometimes it is simply easier to accede to a machine than to exercise power over it. Various perspectives align with this argument. *Cognitive offloading* is defined by Risko and Gilbert (2016, p. 676) as "the use of physical action to alter the information processing requirements of a task so as to reduce cognitive demand." This may be taken quite literally—say, rotating one's head to reorientate an image, rather than rotating the image *in one's head*. Alternatively, 'physical' might be understood in a more object-orientated fashion—say, using an online diary, or even a physical calendar, to remember appointments, leaving one's mind free to forget (also see Barr *et al.*, 2015). Whereas cognitive offloading concerns the *substitution* of cognition through physical actions or external means, *exogenous cognition* is "the technological and algorithmic *extension* of (and annexation of) cognition" (Smith *et al.*, 2020, p. 53, emphasis added). If one used a recommendation algorithm to sort options, before applying one's cognitive resources to consider only those algorithmically determined to be the best, there would be a net *gain*

Luddites, though crucially, modern fears of automation often focus on earnings and employment, while an essential part of the Luddite movement was about the preservation of knowledge and skills *as well as* economic security.

By contrast, some cultural psychologists have argued that human culture can be understood as a shared, if diffuse, body of collective knowledge—a *collective brain* (Muthukrishna and Henrich, 2016). That we can each survive while possessing only a fraction of the knowledge or skills needed to survive is evidence of a successful society and may even be a definition of civilisation (Muthukrishna, 2023). Though, the objection of social critics such as Illich (1973) is not that we are interdependent with one another, but that people are rarely free to *choose* the conditions of their interdependence.

in cognition, as cognitive effort is saved in the initial sorting (Simon, 1987a).[215]

Both have links to notions of *extended mind* (Clark and Chalmers, 1998) and *external cognitive systems* (Barr *et al.*, 2015) whereby technologies and objects in the world are seen as components of, and companions to, human thought. Frischmann and Selinger (2018, p. 81) offer the alternative term, "cognitive prosthetics," but as Smith *et al.* (2020) note, technologies such as AI systems, recommendation algorithms, and smartphones are rarely passive *replacements* for human thought (i.e., a prosthetic), but rather, different means of managing and analysing information which may change how people *actually think*.[216]

At their most extreme, all these perspectives imagine scenarios where people *outsource* decision-making to technologies like AI systems. Sunstein (2024) has made an interesting argument that, often, allowing an AI system to automatically take a decision on one's behalf is worthwhile. Some decisions are time-consuming and boring, and many are routine (Sunstein, 2014). People may find themselves better able to pursue that which matters to them if many decisions are automatically attended to by an AI system, or if many of the decisions they face are automatically architected, via autonomous choice architecture, to encourage them towards a particular outcome.[217] Thus, some might not overrule an

[215] Recent research into the effects of generative AI as a tool for cognitive offloading in education reveal an important twist to this somewhat benign story. Emerging research suggests that generative AI reduces the cognitive burden involved in some educational tasks, such as researching and report writing tasks. However, evidence also suggests that generative AI reduces the criticality and depth of argument from participants who cognitively offload to generative AI, compared to those participants who do not (Gerlich, 2025; Stadler *et al.*, 2024; Valcea and Hamdani, 2024).

Insofar as criticality and depth of argument are often the skills one actually values (e.g., Acar and van den Ende, 2016), this may be a substantial negative consequence of using AI technologies. Furthermore, that AI technologies *do* reduce cognitive burden is largely irrelevant if these 'saved' resources are not effectively 'reinvested.' As with most resource saving technology, it matters little that a technology can save one resources. What matters is *what is done with what is saved.*

[216] The classic thought experiment surrounding extended mind is that of the notebook. If one writes down directions in a notebook, one can forget those instructions and still successfully navigate as the notebook 'thinks' for the person. This is specifically a kind of exogenous memory, an idea which has long been noted in terms of how organisations 'remember' (Simon, 1997a). As Smith *et al.* (2020) note, though, there is a fundamental difference between 'dumb' tools like a notebook and 'smart' machines like an AI system. The latter acts autonomously and is 'smart' through functionality such as feedback and real-time updating. Notebooks, though, do not change.

[217] In discussing authority in organisations, Simon (1997a) also notes that in many instances, people do not follow orders because they feel *compelled* to, but because they consider it *convenient* to. That there are many decisions people do not wish to make ('choosing not to choose', as Sunstein puts it) is an area of behavioural science which has received relatively little commentary, though is of

autonomous system because they recognise the advantages of outsourcing a decision to that system or allowing that system to architect choices for them to reduce the cognitive burden of deciding.

The second reason is *automation bias*. Following from Chapter 4, automation bias is a form of artificing, or human-AI interaction. It describes the tendency for people to follow automated systems, even when such systems should not be followed. This might be because there is a known error, or because the artificing individual has information external to the system which should suggest an alternative approach. In the case of a doctor, automation bias might involve not testing a patient *because* an automated system suggested as such, despite reasons *to* perform the test.[218] In the case of a behavioural scientist, automation bias might involve allowing autonomous choice architecture to nudge a person towards *x* *because* the system determined such an intervention, despite the behavioural scientist having reasons to suspect that *x* may not be desirable.

As in Chapter 4, though, evidence for automation bias is weak. Or, at least, automation bias is more complex than perhaps the word 'bias' would imply to a behavioural scientist. Early studies into automation bias focused on automatic decision aids in practical situations, such as training aircraft pilots. Skitka *et al.* (1999), for instance, found that those trained on flight simulators which used automatic error and failure event notifications performed worse than counterparts trained without these notifications, during failure events (in simulations) where these notifications were not available. This is known as an error of *omission*—missing something because one is not used to spotting it. Skitka and colleagues also found that those in the automatic group exhibited errors of *commission*—doing something simply because it was automatically recommended. Quite dramatically, the researchers found that those in the automatic group had a higher

paramount importance within many normative discussions of when, if, and how one ought to influence another.

[218] Consider the following thought experiment. A doctor correctly diagnoses patients 70% of the time. An AI algorithm is better, with a 90% accuracy. Most of the time, doctor and algorithm agree, and on average, the doctor follows the algorithm's recommendation. Say a patient comes into the hospital, and the doctor is *sure* they should treat for disease *x*. However, the AI algorithm recommends treating for disease *y*. What will the likely outcome be? One might speculate that the doctor will treat for disease *y*, *despite* what they actually think. This is not a matter of accuracy—the doctor is *more confident* they are right than their accuracy reflects. Rather, it is a matter of *accountability*. If the doctor follows the algorithm, and it is wrong, no one will blame the doctor. But the reverse requires the doctor to 'put their neck out' and take on the blame if they are wrong (which they still might be). This is not easy to do. Such a scenario invokes Davies' (2024) notion of an *accountability sink*, where deference to automated systems is done to avoid personal accountability for decisions.

tendency to follow automatic notifications, even when the recommendations of these notifications contradicted other information available to the participants.

By contrast, Alon-Barkat and Busuioc (2023) find no evidence of automation bias in their study of AI recommendations for policymakers (though, per Chapter 4, they *do* find evidence of selective adherence). Two systematic reviews of the automation bias literature document a multitude of factors which colour this complicated scene. Goddard *et al.* (2011) argue that the level of decision-maker experience; the decision-maker's confidence and trust in the system; the type of task being examined; and individual differences in decision-making style all appear to influence automation bias. Alon-Barkat and Busuioc, for instance, suggest that a recent, high-profile algorithm scandal may have influenced the behaviour of participants in their study. Participants were policymakers from the country in which the scandal had occurred. That these participants had reason to mistrust algorithmic recommendations may explain the observed lack of automation bias.

Lyell and Coiera (2017) argue that task verification complexity is a key factor in automation bias. Verification complexity is how easy or difficult it is to assess a recommendation. In the area the researchers focus on—healthcare—it can be extremely difficult to verify whether a test should, or should not, be performed. In such a scenario, verifying a recommendation can be cognitively taxing, and divert cognitive resources from other important matters—say, treating a patient. Thus, in such situations, medical professionals may fail to overrule an automated system. Note that this is not necessarily comparable to above discussions around extended mind and cognitive offloading. Those perspectives typically imply a degree of *choice*; that technologies are deliberately deferred to so as to reduce cognitive effort or improve decision-making outcomes. In instances of verification complexity, decision-makers are *compelled* to defer to the automated system, owing to the constraints on their cognition created by their environment (Cummings, 2015; Simon, 1981).

In a follow-up study, Skitka *et al.* (2000) asked an interesting question: does accountability influence automation bias? The researchers tasked participants to engage in a similar flight simulation exercise as used by Skitka *et al.* (1999). All participants received automated prompts which, as above, sometimes gave recommendations which contradicted other information available to the participants. A control group was informed that their simulation was not being recorded, and that no follow up questions would be asked. A treatment group was told a recording was being made, and that in a follow up meeting, they would be asked to justify their decisions in the simulation to a researcher. The researchers found that those held accountable made fewer errors of *omission*, spotting more vital details when no notification was given, compared to the unaccountable control group. However, the accountable group made more errors of *commission*, responding to notifications even when other evidence contradicted the recommendation.[219]

This result is not necessarily surprising—those who were held accountable were more attentive to *everything*, noticing that which they were not prompted to notice, and *definitely* noticing that which they were prompted to.[220] Yet, this result also adds further layers of complexity to automation bias. Accountability may both *increase* and *decrease* automation bias, and as an intervention, its effectiveness depends on whether one cares about spotting things which might be missed (omission) or avoiding things which should be avoided (commission).[221] It is such a fractured landscape that leads Goddard *et al.* (2011, p. 125) to conclude that, "there are enough studies, discussion papers, and anecdotal evidence to imply that it [automation bias] is a consistent effect," but the "major unresolved issue is the incidental nature of the reporting of automation bias."

Still, perhaps something can be taken from all this. Automation bias, rather than being an inherent human behaviour, appears to be much more contextually derived. One's relationship with technology, and the subject matter; the nature of the task itself; and the environment in which the task occurs; all seem to be factors which influence whether people demonstrate discretionary power and overrule autonomous systems.[222] Such a conclusion is not necessarily surprising (e.g., Simon, 1997a). Yet, it does

[219] Given that commission is more indicative of automation bias (i.e., *following* an automated system) than omission is, this difference is likely important.

[220] As above, it is quite reasonable that the very act of prompting 'leaked' information to participants that that thing was important.

[221] In relation to AI specifically, an interesting hypothesis is that of *expectation bias*. Broadly, by virtue of calling the technology *artificial intelligence*, people approach the technology with an expectation of its high intellect. As in Chapter 4, intelligence arises through social construction. If one expects an AI system to produce an intelligent, useful, and *superior* answer, one may readily deceive oneself, warping the facts of the matter through *ex post* rationalisation to make the output of the AI system align. This might also have interfaces with information leakage.

[222] One might draw parallel to the persuasion-knowledge model discussed in Chapter 3. There, one's metacognition is said to be influenced by one's own knowledge of the subject at hand. Where one is ignorant, one is held to be persuadable, and may *automatically* follow the recommendations of, say, a salesperson. As such, both automation bias and metacognition may be understood as probing the conditions under which people discharge (or recharge) their agency.

encourage exploration of a third reason for adherence: that discretion is *nominal*.

In many instances, people do not *actually* have the ability to overrule an automated system. They might *legally*, and also often do *physically*, but the social and political forces which define and shape the institutional context in which one finds oneself ultimately prevent a person from exercising this authority. In many ways, then, the third reason is a restatement of various arguments in Chapter 4. But it is, nevertheless, important to restate the point that the nominal authority one has within a formal hierarchy or structure only ever partially (if at all) translates to *actual* authority within the informal heterarchy in which decisions are taken (Simon, 1997a).[223]

Still, there are some additional points to be made, besides blandly stating: reread Chapter 4. Most importantly, that adherence to technology is often because technology is the only means of achieving broad objectives. One cannot get to the moon without a rocket, just as one cannot cheaply produce millions of cars without automated production lines or deliver personalised nudges to millions of citizens without some degree of automated architecting of choices divorced of perfect human oversight. Whether these are ends society should desire is a normative question, or a question of values; accepting that these ends are desired, that technologies *must* be adhered to so as to achieve these ends is a question of facts.[224]

These three reasons—convenience, automation bias, and material circumstance—suggest that autonomy and motive power are not necessarily *absolute* categories but more *states of being*. A tool can, in effect, be a machine so long as the person who gives the tool motive power *always does so*. If the exercise of human discretion is *nominal*, for any of the reasons above, it will often be worthwhile to treat tools *as if* they were machines (*quasi*-machines) possessive of 'their' own motive power (*quasi*-motive power). This is useful to

[223] Writing on US nuclear war plans in the 1960s, Ellsberg (2017) notes that too many US deployment systems relied on individual judgement and discretion as a failsafe, while the military apparatus explicitly *discourages* defiance of senior orders. Thus, Ellsberg argues, no effective failsafe can rely on the discretion of military personnel because, for most, such discretion is *nominal*. As such, Ellsberg reports advocating for superior communication channels and stand-down procedures—essentially, means by which subordinates could be ordered to abandon an 'automatic' attack plan.

[224] Though, the *degree* to which technology is used, and *how* technology is used, remain outstanding questions. As Chapter 2 outlines, one can personalise nudges using limited data, and thus through a simpler automatous system that more immediately preserves accountability and explainability. As Chapter 4 outlines, *that* technology will be used does not close off the question of *how* technology should be used. Societies that desire ends which require technologies are thus not slaves to technologies but retain (and must *entertain*) many questions of about norms, values, and desired futures.

recognise in relation to algorithms to support decision-making, as discussed in Chapter 4.

Yet, this chapter is about autonomous systems which do possess motive power. To this end, the above reasoning for a lack of human discretion—which remains, simply in a different form—is relevant insofar as it shapes what people think automated systems *ought* to do. Once the input of a doctor is turned over to an automated system, the question shifts from *how should this patient be treated?* to *what treatment should this patient receive*? For a behavioural scientist, the same shift arises, from *how should this person be nudged?* to *what nudge should this person receive*?[225]

One should anticipate such a shift, as one is seeking to replace a social interaction—between a doctor and patient, or a choice architect and decision-maker—with an automated interaction—between a machine and a person. The machine must be given a logic for operation, and a set of parameters by which it can act upon the world.[226] There are rules of the game which machines obey.[227] People have more flexibility, discretion, and perception which allow them to operate with looser parameters and rules (Simon, 1997a). While this discretion leads some to advocate for the introduction of algorithms in decision-making, this same discretion supports arguments—by those same advocates—for the retention of people all the same (Sunstein, 2024, 2023).[228]

[225] A worthwhile example is that of the car. Before the car, communities were defined by one's ability to walk somewhere; necessary services had to be sufficiently numerous as to accommodate these human limits. After the car, communities were defined by one's ability to drive somewhere. This changed how provision was undertaken, and the questions contained therein. Rather than asking, as before, *is this community provisioned with this amenity*, the provision is taken for granted, and matters turn to questions such as *can this road support an adequate level of traffic* or *are there enough routes linking enough places*? Matters are no longer about increasing provision but increasing *access* to existing provision.

Human efforts have, today, shifted away from providing services we need, to providing the conditions for the technologies which provide those services (Illich, 1973). Whether this is a better or worse situation than before is a matter of opinion. But that cars (and many other technologies) have changed the choices people must make (via the questions one must ask) and the assumptions that are made cannot be discounted. Neither can one ignore the behavioural outcomes of these changing choices.

[226] Just because a machine can act upon the world does not mean it can act *in every way* upon the world.

[227] Problems arise not through machines 'breaking the rules' but through the rules being poorly specified by people. This is the essential foundation of Bostrom's (2014) famous 'paperclip problem' where ill-defined rules could cause an autonomous system tasked with manufacturing paperclips to destroy the world.

[228] One might wish to reconsider the notion of technological disintermediation from Chapter 4. The introduction of technology which displaces people also changes how discretion can be exercised. A bank teller might be able to advise a

Therefore, when dealing with autonomous systems, one is not dealing with systems and processes in which people are absent, but in which social relations have changed through the necessary introduction of new assumptions and logics about the people involved.

Being Artificial

Autonomous choice architecture may be synonymous with targeted nudging or adaptive nudging (Mills, 2022b; Peer and Mills, 2024). While *in principle* these forms of personalisation may be achieved through manual, rather than autonomous, means, in practice, an autonomous infrastructure is essential. For both, this is because of the extensive amount of data which must be handled to target the intervention. For adaptive nudging, that real-time monitoring and updating may be used within the personalisation process also strains—to an insurmountable degree—the means by which this approach could be undertaken manually. In most instances, and essentially in *all* instances in the case of adaptive nudging, an autonomous system will be required, and thus autonomous choice architecture will be used.

These forms of personalisation are the starting point, though one could consider autonomous systems with more capabilities. For instance, targeted nudging and adaptive nudging *could* have a finite set of choice architecture features and a static decision-tree based algorithm to determine the design of the intervention. *Or*, one might imagine a system which incorporates elements of generative design, dispensing with the set of features and creating bespoke features in accordance with (and in response to) decision-makers. Additionally, one could incorporate a predictive AI component and personalise the *outcome* as well as the design. To date, flavours of all these systems can be found, most commonly recommendation algorithms which incorporate some kind of visual personalisation.[229]

In response to these systems of autonomous choice architecture, behavioural scientists—or, more commonly, technology critics of behavioural science—have proposed various new terms.

Firstly, rather than talk of choice architecture, it may be increasingly helpful to talk of a *choice environment* (Mills, 2022b;

customer on the best savings product—an act of discretion likely within the tolerances of the bank. Banking apps, typically, do not.

[229] As noted in Chapter 2, Netflix is reported to personalise the recommendations given to viewers, but also the thumbnails based on what it predicts will entice a viewer. Also, as above, the use of generative design or website morphing techniques is more speculation than fact. Though, these attenuations are natural extensions of many of the ideas discussed here.

Yeung, 2017), as was done in Chapter 2. Assuming a high degree of malleability, many aspects of the choice architecture which surround a decision may be personalised by an AI system to influence a person's choices. Furthermore, as discussed below, such systems often seek to influence people over multiple decisions, or *over time*. Thus, when considering such a system, one might question whether architecture is too narrow of a term. An architect designs a building through which a person moves, and the layout of the building influences (without always forcing) how a person moves through it. But the architect does not design the buildings *next* to that which is in their gift. A city planner, though, may have the power to wholly redesign the placement of buildings, and the flow of people to and from those buildings. As malleability increases, it may often make more sense to talk of designing a decision-maker's *environment*, rather than merely the *architecture* of choices.[230]

Secondly, there is the concept of potency (Yeung, 2017). Potency was alluded to in Chapter 2, though is more helpfully addressed in relation to autonomous choice architecture.[231] When one uses an impersonal nudge, the effectiveness of the nudge can be measured in terms of *how many people choose the option advocated by the nudge*. One faces a curious problem when nudges are personalised, though. Assuming that everyone is still nudged towards the same option, the effectiveness of personalisation must then be defined as *how many more people choose the option advocated by the nudge*. Assuming that the option advocated for was also personalised (e.g., through personalised paternalism), the effectiveness of the personalisation is actually given by *how many more people choose the option advocated by the nudge, whatever that was*. In both instances, personalisation is about *increasing how many people do as they are nudged to do*. From this perspective, a nudge's potency may

[230] In most cases, malleability is very high when dealing with autonomous choice architecture. This is for two reasons. Firstly, to make it worthwhile to use an automated system, one must typically be dealing with large amounts of data and many different intervention designs. This is to say, malleability creates the need for an autonomous system, and so if one has an autonomous system, malleability will also, often, be high. Secondly, scaling a behavioural intervention is one of the major benefits of an autonomous system. This, again, might be a chicken-and-egg situation—if one has an autonomous system, one likely wants to achieve scale. Regardless, the point still stands that scale may be important. Scaling an intervention raises once more the problem of heterogeneity and makes the use of personalisation more compelling (Mills and Whittle, 2024a; Sunstein, 2022a). Thus, malleability, scaling, and autonomous choice architecture often form a loose triptych, and a reader may benefit from holding all these concepts in their mind henceforth.

[231] The term 'potency' was originally used by Yeung (2017), though this term—and others—were not fully defined in that work. The perspective discussed here draws generally from Mills (2022b) and is the interpretation of potency which is both most sensible for this discussion and most amenable, in my opinion, to how Yeung uses it.

be defined as the percentage of people who choose the option they are nudged towards (Mills, 2022b). While a successful nudge, paradoxically, seeks to have some people *not* follow it—as a demonstration that people still can 'go their own way' (Thaler and Sunstein, 2008)—successful personalisation seeks to have a high potency, and thus as many people follow the nudge (Frischmann and Selinger, 2018).[232]

Thirdly, there is *hypernudge*. Yeung (2017) introduces the term hypernudge rather loosely to describe how adaptive nudging powered by autonomous choice architecture threatens privacy and dignity in society. Loose definition has caused the term to have a troubled life. Frischmann and Selinger (2018) do not use it, but frequently describe instances which might be likened to hypernudging in their description of techno-social engineering. Darmody and Zwick (2020) *do* use the term hypernudge, but as part of a wider discussion of consumer influence as a strategy in the modern digital economy. As hypernudge has garnered more attention, it has entered the lexicon of behavioural science (to an extent) as a kind of byword for a digital nudge, or a technologically mediated nudge. All conflate the term more than they elucidate.

Morozovaite (2021, 2020) has undertaken admirable and interesting work defining features of hypernudging and relating it to wider regulatory frameworks. For instance, Morozovaite (2020, p. 118) notes that hypernudging must comply with features of behavioural nudging—changing only choice architecture (not economic incentives), leaving options open (no bans or mandates), and utilising behavioural insights. Yet, hypernudges must also a) be delivered through digital interfaces; b) be personalised; and c) change in response to d) predictions about people. Taking Morozovaite's (2020) analysis alone, one can interpret hypernudging as an alternative name for adaptive nudging (Peer and Mills, 2024). Might one thus choose to ignore the term, or select language based on the domain of interest—adaptive nudging for the behavioural scientist, hypernudging for the regulator or critic.

Such a compromise would not be unreasonable.[233] Though, this does not mean the idea—whether dubbed hypernudging or

[232] This tension, to my knowledge, has not been substantially explored within behavioural science, but it is at the core of many critiques of behavioural science as techno-social engineering, or as a social technology (e.g., Frischmann and Selinger, 2018; Yeung, 2017). This tension is the catalyst for much of what is discussed in this section.

[233] In my time writing about hypernudging (Mills, 2022b), and as one of the originators of the term 'adaptive nudging' (Peer and Mills, 2024), I have sometimes found the term 'hypernudge' to be a bit of an unnecessary confusion within discussions. I have often been asked (by behavioural scientists) to give an example of a hypernudge, or to explore what could be done to a nudge to make

adaptive nudging—does not benefit from further exploration, as Mills (2022b) has attempted to do. The argument is that hypernudging should not be understood as a 'type of nudge' but rather a *system* of nudges; just as *hyper*text connects webpages, and *hyper*space connects star systems (at least in science fiction), *hyper*nudging is a system of nudges connected by a technological infrastructure. This infrastructure consists of those features described by Morozovaite; specifically, a) data about decision-makers, including real-time adherence data; b) predictive algorithms (including AI systems) to analyse these data; and c) digital interfaces with adequate malleability to personalise interventions in response to predictions.[234]

Mills then asks whether this system, and the logics contained within it, are compatible with (analogue) behavioural science. From the perspective of real-time feedback, it may be tenuous at best to suggest they are compatible. Consider the case of automatic enrolment in organ donation—a classic nudge example (Johnson and Goldstein, 2003). In an 'analogue' setting, one will experience this nudge only periodically—say, every five or ten years when they renew their driving licence. Were one to reject the nudge, and opt-out, there would not be much more to it. One *might* object to them being nudged in the future (at renewal), but equally, one might argue that provided adequate time has passed, it is reasonable to speculate that one's preferences may have changed, and that nudging again is acceptable.[235]

The question of *how much time should one wait between nudges* cannot be easily answered. But there is certainly a worthwhile argument that waiting *some* period of time after a nudge has been rejected is fundamental to the premise that a choice architect respects one's decision (Lades and Delaney, 2022). Indeed, while rarely considered 'coercive,' repetitive or endured asking may at

it hyper, and so on. As a reader will see, such questions cannot be answered when hypernudging is understood as a *system*, as I (Mills, 2022b) define it. To these questions, adaptive nudging is better suited, precisely because of its taxonomical description and definite constraints (e.g., data and malleability). From my experience, critics and those more aligned with the techno-social engineering or social technology perspective have had an easier time using the idea of hypernudging, in part because many of these discussions and critiques do not centre on specific intervention designs (which is not necessarily any more helpful than the questions sometimes raised by behavioural scientists; again, see Mills, 2022b).

[234] The idea of a system of nudges is why Peer and Mills (2024) argue that, at a certain level of personalisation, it no longer makes sense to talk of 'nudges' and that instead one should talk of 'nudging' as a process. Hence, targeted nudging and adaptive nudging, rather than targeted nudges and adaptive nudges.

[235] Indeed, one might argue that *never* trying to influence someone after a decision has been made is worse than nudging periodically, as they may have chosen something they now regret. See, for instance, various subscription services.

times feel as coercive as force, or economic pressure, by virtue of being a hindrance, or of being so *annoying*.[236]

Yet, this is the situation one might experience when interacting with a hypernudging system: adherence data (e.g., that one rejected the nudge) feeds into predictive algorithms *in real-time* to dynamically reconfigure the choice environment and *immediately* nudge the person again (Mills, 2022b). In some instances, the 'ask' of the nudge might be slightly different—like asking for *just* one's kidneys, rather than all of one's organs—but the principle is the same. A person has made a choice; it is not in accordance with the outcome desired by the system; and so, the system intervenes again to achieve a 'better' outcome. To this end, Mills (2022b, p. 5, original emphasis) succinctly argues that "hypernudges *follow*."

It would be difficult for a behavioural scientist to argue that, under such a scenario, freedom of choice has been preserved. Likewise, it would be difficult to argue that the decision-maker's dignity has been maintained, or that they have been respected. A choice architect can and often will do these things; an autonomous choice architect need not, and often, will not be designed to do so.

Such a state of affairs exists because of the logic which must be applied when designing autonomous choice architecture. A behavioural scientist, whether they realise it or not, understands that their intervention exists within a wider social context. The notion of potency is an alien one in behavioural science not because it is unique to an AI system or some other component of the technological infrastructure of hypernudging, but because it is not an especially *natural* or *intuitive* way of thinking about what a nudge *actually* does. Behavioural scientists, I suspect, implicitly view choice architecture not as steering mindless entities through abstract processes (which can be measured by some measure of 'efficiency' like potency), but as changing how people *experience* information, showing new arguments or reframing choices so that they might evoke new thoughts and feelings (Sunstein, 2017).[237] Whether biased or nudged, people make choices for reasons, and behavioural scientists implicitly recognise this.

It is this recognition that makes potency a bit of a weird idea for analogue behavioural science. Behavioural scientists do not necessarily care how many people followed a nudge—they will

[236] What, in the dark patterns literature, has been called *nagging* (Gray *et al.*, 2018).

[237] Of course, this might be rejected. For instance, what is described above could be taken as suggesting behavioural scientists seek to nudge people to think more deliberatively about their choices, though this is often not the case (Beshears and Kosowsky, 2021). Furthermore, there *is* a compelling argument to be made that the suite of behavioural biases which underpin nudges implicitly leads to the treatment of people as 'mindless entities' even if this is neither the intention nor admission of behavioural scientists

often care more about evidence that a nudge had an effect, or that people choose options they do not come to regret. But an autonomous choice architect lacks this context. Adherence is a yes/no variable, and necessarily, the system will be engineered to maximise yes's (*yes, they followed the nudge*) and minimise no's (*no, they rejected the nudge*). From the perspective of a computer scientist or data scientist, viewing adherence as a yes/no variable, the idea of potency is much more intuitive. It is also necessary as a variable for a predictive algorithm to *maximise* (Russell, 2019).[238]

There is thus a compelling basis to claim contradiction between the ideals of behavioural science and the realities of autonomous choice architecture—to achieve the latter, one must abandon elements of dignity and respect which are valued by the former.[239]

Though, one might raise an objection with a line given above, where it is noted that a hypernudging system might change what a person is nudged towards to raise potency—asking for *just* one's kidneys, not all of one's organs. If such adaptive personalisation leads the system to eventually nudge a person towards that which they would already have chosen, is there really a substantial issue here? To this question, one might respond: *yes!* Indeed, one can raise objections methodologically and ethically.

Methodologically, if the system eventually nudges towards an outcome the decision-maker could have chosen themselves, the system is not especially useful. This is a variation on the selection bias problem discussed in Chapter 2. To avoid this problem—to have a reason for the system to exist—one must assume that the decision-maker *cannot* identify the outcome they ideally desire. Behavioural science makes this assumption by appealing to the suite of behavioural biases identified over several decades (Thaler and Sunstein, 2003). To this end, a hypernudging system could be

[238] One might add to the list of computer scientist and data scientist the politician or the policymaker. Insofar as these groups may care more about the appearance of a policy being successful, rather than the policy leaving people better off *as judged by themselves*, a potency perspective may be appealing. It is interesting to note that Yeung (2017), in the first use of the term 'hypernudge' discussed the concept in relation to a unique form of *regulation*: regulation *by design*. One might interpret this language as meaning the use of design-based methods (rather than economic methods or force) to achieve *compliance*.

[239] Darmody and Zwick (2020) pick up on this idea, though not directly from a behavioural science perspective. They argue that notions of empowering consumers, through proliferation of choice, appear in contradiction if, at the same time, digital surveillance and autonomous choice architecture are utilised to constantly manipulate consumer preferences. Consumers, thus, are not empowered. Though, they might *think* that they are—a potentially interesting inflection on the idea of machine laundering. Or it might contribute to a whole new idea—*nudge-washing,* or the use of nudges to claim choice where none practically exists.

said to do the same, nudging precisely because people are biased in their decision-making.

Thus enters the ethical objection. All choice architects (manual and autonomous) begin with an assumption that people are biased. This is a major point of contention for some in the field (e.g., Gigerenzer, 2015) but is not wholly objectionable insofar as behavioural scientists accept that people might choose to reject a nudge, as above. Rejection does not necessarily change whether a behavioural scientist believes someone to be biased. But in respecting a person's choice—whatever that might be—the behavioural scientist accepts that their beliefs do not justify further intervention.

However, hypernudging systems never stop nudging people. They are imagined as continuously reconfiguring interventions in response to what people do, constantly seeking to steer people towards particular options. A recommendation algorithm, for instance, never *stops* recommending things to you. Hypernudges, in essence, codify assumptions around the fallibility of human judgement, and proceed on the basis that people *always* need help in choosing. A name can be given to this design characteristic: the *assumption of error* (Mills, 2022b). This is to say, autonomous choice architecture assumes people to *always* be in error; to *always* need a nudge. That the system might, eventually, nudge someone in a desirable way does not erase this fundamental aspect of the system's design.

The assumption of error is interesting insofar as it fits into a longer, more substantial critique of how technology and language have evolved to change how human behaviour is understood, and thus what interventions may be 'acceptable.' For instance, Skidelsky and Skidelsky (2012, p. 155) argue that, today, 'healthy' is rarely taken as a qualitative statement aligning with "everything working as it should" but has instead evolved into a quantitative assessment that encourages a view of health as something subject to "perpetual improvement." By recasting healthiness as something to be measured, it recasts anything less than maximal as 'unhealthy,' justifying some medical intervention.

One can perhaps see how the assumption of error emerges from quantification by considering the work of Gilbert S. Daniels. Daniels (1952) was hired by the U.S. Airforce to design a one-size-fits-all jumpsuit for pilots.[240] He and his team took dozens of physical measurements of pilots to determine the average specifications from which to base the design of this new jumpsuit. However, Gilbert soon discovered an issue. The more average measurements he combined together (e.g., average leg length and

[240] Diligent readers will recall this example from a footnote in Chapter 2.

average arm length), the fewer pilots actually conformed to the average. Very quickly, *no one* was average—everyone demonstrated *some* deviation from the desired specifications. The truly average jumpsuit thus became a one-size-fits-*no-one* jumpsuit.

This example demonstrates that with more data and data analysis, it becomes frightfully easy to find *something* where a person deviates in some degree from some 'desirable' or 'preferential' metric. To address Skidelsky and Skidelsky (2012), it is no surprise that modern conceptions of health are more about "improvement" than about the body "working as it should." Simply by measuring age, weight, blood pressure, heart rate, alcohol consumption, and so on, a sufficiently powerful suite of data analysis tools could identify *something* which a person might 'improve' upon. As Skidelsky and Skidelsky (2012, p. 156) write: "[I]f there is no such thing as perfect health, then *any* undesirable condition can be defined as illness and made an object of medical treatment. If every state of the body can be seen as defective relative to some other, preferred state, then we are all in a sense perpetually ill."

The social critic Ivan Illich (1978, p. 38) introduces the term "disabling professions" to describe areas of life, such as health, education, work, and so on, where greater insights and capabilities do not alleviate people's problems, but actually *create more problems*.[241] As with Skidelsky and Skidelsky, health is a major point of critique for Illich, who argues that modern medicine rarely 'makes people healthy,' but rather, more typically, acts to more specifically define people's ills, and thus prescribe treatments for them. He (Illich, 1973) notes—as does Feyerabend (1978)—that this creates the perception of crisis and emergency in health, encouraging more funds and brains towards solving such issues. Though, inevitably, this leads only to *more* data being collected, *more* insights gleamed, and so *more* problems found. Hence Illich's notion of a *disabling* profession—a profession (e.g., doctor, teacher, manager) tasked with identifying problems and prescribing correctives.[242]

[241] Through a reading of Illich, one would be hard pressed *not* to label behavioural science as a disabling profession. Indeed, much of the immediate analysis pulls on this thread.

[242] Illich (1978, p. 44): "Today, industrial societies are constantly and totally mobilized; they are organized for constant public emergencies; they are shot through with variegated strategies in all sectors; the battlefields of health, education, welfare, and affirmative equality are strewn with victims and covered with ruins; citizens' liberties are continually suspended for campaigns against ever newly discovered evils; each year new frontier dwellers are discovered who must be protected against or cured of some new disease, some previously unknown ignorance. The basic needs that are shaped and imputed by all professional agencies are needs for defence against evils."

Finally, one can turn to Simon (1997a). While perhaps not as radical as Illich or Feyerabend (or even Skidelsky and Skidelsky),[243] Simon too offers insights into this problem. Simon (1997a, p. 256-257), writing in 1947, argues that "the term "efficiency" has acquired during the past generation connotations which associate it with a mechanistic, profit-directed, stop-watch theory." Yet, he continues that "until practically the end of the nineteenth century, the terms "efficiency" and "effectiveness" were considered almost synonymous. The Oxford Dictionary defines "efficiency" [as] "Fitness or power to accomplish, or success in accomplishing, the purpose intended; adequate power, effectiveness, efficacy."" Thus, for Simon, there also has been a transition in perspective, as efficiency has shifted from something more qualitative in character (e.g., adequate, good enough, satisficing) to something quantitative in character, and thus capable of being maximised.

Simon suggests this is why there has been an explosion in data collection, and a pressing need for information management capabilities, without these data or capabilities resulting in substantial benefits for organisations or people. If the demand of a process is that it is 'efficient' (i.e., maximising) rather than merely being 'effective' (i.e., satisficing), then one necessarily needs to measure everything to ensure efficiency is achieved. Yet, as above, more measurement is likely to reveal more problems, demanding more intervention. Even if it does not reveal problems, that *anything* other than efficient is considered undesirable shifts perspectives on

[243] Illich never identified himself as an anarchist, but his ideas fall within that intellectual tradition, and he was acknowledged by contemporary radicals (e.g., Marcuse) for his radical critique of daily life. Feyerabend *did* identify as an anarchist, though specifically an epistemological anarchist, and not necessarily a political anarchist. This did not stop Feyerabend advocating for a society where science was held in the same regard as alchemy or theology or paganism, and so on. He even, at times, made interesting arguments around the superiority of traditional medicine over modern medicine.

Skidelsky and Skidelsky are not radicals of the anarchist tradition and would not identify themselves as such. Still, their work hinges on the argument that the modern world has forgotten the ends it is trying to achieve, and has become obsessed by the means, instead. Hence, rather than thinking about what it means to live a healthy life, they suggest society today obsesses over the classification and treatment of ill-health. In this respect, their critique falls within the same ballpark as Illich's.

Simon (1997a, p. 249) must be separated from the rest not merely due to the lack of engagement with radical politics in Simon's writing, but due to Simon's very explicit endorsement that new technologies and data insights be used to solve problems and improve people's lives: "The new problems created (or made visible) by our new scientific knowledge are symptoms of progress, not omens of doom. They demonstrate that we now possess the analytic tools that are basic to understanding our problems—basic to understanding the human condition. Of course, to understand problems is not necessarily to solve them. But it is the essential first step. The new information technology that we are creating enables us to take that step." This is strikingly similar to a recent argument by Buyalskaya *et al.* (2021) arguing that new data and data analysis tools are ushering in a 'golden age' of social science.

intervention from 'intervene when there is a problem' to 'always intervene.'

One can readily understand the problem of heterogeneity in these terms. An impersonal intervention might be effective, but it is unlikely to be efficient because of individual differences. Only through more data, and personalisation, can an impersonal, but effective, nudge be transformed into a personalised, and *more* effective, intervention, with the (efficient) end being total adherence to an intervention (potency) through ever-more data and ever-more personalisation of an ever-expanding choice environment.

What this detour demonstrates is that the assumption of error is an outgrowth of how human behaviour is understood by the behavioural scientist and how this understanding comes to be translated to a machine. This is a multi-faceted problem. One facet is the prevalence of data—more data increases the likelihood of *some* 'error' in decision-making being found. Another facet is the role of potency—by introducing a component which a machine can maximise, intervention becomes the default action to be taken. But a third facet is the notion of bias itself. By arguing that people are often biased, and thus make errors, behavioural science creates a justification for *some* intervention. This is the starting pistol for the other two facets.

Hence, behavioural scientists must be cautious in using bias as a justification for introducing AI systems and other algorithms into decision-making. Once one accepts that decision-makers may be biased, and that these biases can be eliminated, one may go looking for them—indeed, this has been proposed as a use of AI technologies in Chapter 3. With enough data, one will inevitably identify *something*; and with this identification, one can justify the introduction of a *corrective* in the form of an algorithm.[244] But this logic has no natural stopping point. For instance, one could investigate whether decision-makers are biased in their adherence to an AI system's recommendation, *as has been done* (see Chapter 4).[245] If bias is identified here*, as has been found*, one might argue that people should exercise discretion over an AI system only when *another* AI system 'thinks' they should, and so on.

[244] This is why behavioural science is vulnerable to being used as a tool for machine laundering. For someone who *already* wants to use an algorithm, arguments that people are biased are a compelling way of transforming this political objective into an 'objective' fact. See Chapter 4.

[245] Also see an interesting, recent paper by Glickman and Sharot (2024), which finds that AI systems often reinforce human decisional biases, leading to greater bias than had people acted *without* the AI system. Such a finding could be used to justify the elimination or minimisation of human involvement *at all.*

A science-fiction fan may wonder if this is an argument for machines taking over. It is not. Machines are not autonomous insofar as they may choose to overthrow humanity. But machines are sufficiently autonomous that people may choose to use them to control, influence, and displace other people. It is to this end that behavioural science can be described as a *social technology*. In discussing the history of money, Martin (2015) defines a social technology as, "a set of ideas and practices which organise what we produce and consume, and the way we live together."[246] Behavioural science, and ideas such as behavioural biases, can be understood on similar terms. It is not that machines are taking over; following behavioural science as a social technology, it is that people are fallible, and this may justify the introduction of a machine. Ideas like biases organise how we think about one another, and the responsibilities we give to one another.

But social technologies are not perfect. Money exists, but so too does charity, and community governed through non-monetary relations.[247] Any frequent rider of a city bus will attest that people, on occasion, may be waved on despite not having sufficient funds. You are expected to pay for the bus, though you might not be able to, and so people (a driver) step in to keep things going. As above, behavioural scientists think about people as being biased, and that these biases may lead people to make choices they come to regret. But these are organising principles that are not perfect, leading to provisions such as the need to let people *go their own way*.

However, when social technologies are transformed into autonomous technologies, these imperfections create tensions and form the basis of objection. This transform happens in two ways. Firstly, the language and practices of the social technology form the basis for developing the autonomous technology. For instance, digital interfaces are transformed into choice environments consisting of different architectural components which may be changed; behaviours are transformed into some variable called 'adherence' which, in the aggregate, is called 'potency.'[248] But secondly, social technologies establish the *rationale* for the introduction of autonomous technology. Once people make 'biased' decisions, there is a rationale for the introduction of 'unbiased' technologies to 'support' decision-makers through prediction (e.g., hypernudging, autonomous choice architecture) or

[246] My decision to use the term *social technology* was made prior to discovering Martin's definition. Though, this definition is succinct and appropriate—far better than I might define the idea myself (see Chapter 2)—and so I will proceed with it.

[247] Though one may be concerned that such practices are under threat.

[248] For examples involving money, tips become automatic 'service charges,' trust becomes a 'credit score,' and so on.

recommendation (e.g., algorithms to support professional decision-makers).[249]

The Impersonal Touch

These are objections. It does not necessarily follow from them that the use of behavioural science as a social technology, or the development and implementation of behavioural technologies grounded in this social technology, should be *rejected*. Nevertheless, as these technologies emerge and continue to develop, faster and faster with the development of AI technologies as 'decision centres' for autonomous systems, it is important to consider these objections, and—perhaps—reject some applications. What, then, are the solutions, or, at least, the alternatives?

In designing the jumpsuit, Gilbert realised that he could not design one which fit everyone. Neither could each pilot receive a tailored jumpsuit—it would have been timely, costly, and undermined the interoperability of military logistics. The solution, in a manner of speaking, was to introduce some waste into the system. Rather than designing one which fit an imaginary, average, pilot, Gilbert recommended a jumpsuit based on average measurements, *plus a bit*. Allowing for some variability and error was inefficient, both in terms of material usage and in terms of the number of pilots equipped with a form-fitting jumpsuit. But it was an *effective* solution because most pilots fell into the category of average, *plus a bit*. By sacrificing optimality, Gilbert was able to actually *solve* the problem of designing a standard issue jumpsuit.[250]

[249] For instance, the 'invention' of value was used to justify the enclosure of land and the transformation of agrarian society into a market-based one (Biss, 2022; Polanyi, 2024). For centuries, land in rural England was held in common, with its 'value' being the sustenance which the land would provide for the peasantry. The land was 'effective,' even if not used *efficiently*. Enclosure (the privatisation of common land) was justified, though, on the basis of *improvement*, or that the peasantry was neglecting the value of the land, and that it needed to be transformed into property to be 'efficiently' used. Thus, the notion of 'value' was used to justify the imposition of technology (private property and the commodification of land).

[250] As Simon (1997a) argues, today, questions of efficiency or optimality are often seen as the problems to be solved. But efficiency or optimality only ever exist within the context of a different, more important problem. That jumpsuits could be comfortable or cheap was desirable, but the actual problem was designing a standard issue jumpsuit. Sacrificing a solution to the overarching problem so as to attain these second-order goals should thus be avoided; nevertheless, Simon (1997a) argues this is sometimes not the case, leading to poor management, worker satisfaction, organisation performance, and so on.

Of course, if the solution to a problem is an unacceptable one (e.g., it is too expensive), there may be a rationale in thinking about efficiency or optimality. But if one has a problem with *no* viable solution, no amount of efficiency is going to help. One needs to either a) reformulate the problem such that a viable solution can be found; or b) reformulate one's goals to bring one's capabilities in line with one's options.

A similar phenomenon is found in organisations, too. As discussed in Chapter 4, managers who have limited attention may often allow workers freedom to shirk work and slack off simply because it is not possible to attend to these undesirable behaviours *and* all the other matters which fall within a manager's remit (Cyert and March, 1992). Thus, the effective solution is to allow some slacking and accept a generally adequate level of worker performance, rather than to pursue the 'efficient' solution of maximising worker performance, which will inevitably be much more difficult, and may be more costly.[251] Sometimes, the 'efficient' solution actually leads to worse outcomes (Friedman, 1987).[252]

These examples may all be described as *satisficing* solutions—accepting a 'good enough' outcome when the 'best' outcome is costly and difficult to achieve. To this end, one might look at the question of personalisation itself and ask whether *impersonal* behavioural science should not yet be written off. Arulsamy and Delaney (2022) investigate the effect of automatic enrolment into workplace pensions in the UK. For years, policymakers and political leaders had been concerned that UK workers were not saving enough for their retirement, hence this policy introduction. Arulsamy and Delaney note that there are various reasons why a person might not save, such as mental health difficulties. This barrier was reflected in savings data—those who suffered from some mental health difficulty saved, on average, less than those who did not.[253]

This, then, is a prime instance where one might think a personalised intervention should be used. There is a clear individual difference (mental health) which could, feasibly, encourage one to see this group (those with mental health difficulties) as benefiting

[251] Note that maximising worker performance is 'efficient' in the sense of getting the most out of one's workers. If the costs of achieving this are inhibitive, as they often are, then this can hardly be called an efficient solution. But here, efficiency would now mean two different things—worker efficiency, versus organisational efficiency. The latter is almost certainly more important than the former, hence why the former will often be abandoned, and an effective solution implemented.

[252] Friedman (1987) critiques Marxist accounts of management, suggesting that these too strenuously adhere to ideas of Taylorist *scientific management* as the actual form of capitalist production. Much like Simon (1997a) has criticised the Taylorist mindset, Friedman argues that 'pure' scientific management (and thus the pursuit of scientific 'efficiency') would create much more conflict than value for an organisation. Instead, capitalist organisations permit some worker discretion (amongst other things) to achieve an effective worker output, one which is sufficiently high while reducing the risk (or severity) of resulting conflict.

[253] The reasons why are quite intuitive. If one is preoccupied with one's mental healthcare, and the implications therein, one will have less time and attention to devote to something dull, like planning for retirement. Conditions like depression, which are *generally* demotivating, are unlikely to correlate with especially proactive retirement saving.

from a personalised intervention. But what is interesting about Arulsamy and Delaney's study is they find an impersonal intervention—automatic enrolment for *all* workers—both *eliminated* the savings rate gap between those with a mental health difficulty, and those without, *and* increased average savings of *all* people. From an efficiency (maximising) perspective, maybe a personalised intervention would have been most effective—it is feasible that some 'problem of heterogeneity' was present here.[254] But from an effectiveness (satisficing) perspective, an impersonal intervention worked.[255]

For all the possibilities of AI technologies enhancing behavioural science through techniques such as personalisation, this chapter has sought to articulate various social and ethical costs which arise when these technologies are utilised. The chapter has not even begun to touch on a litany of others. Let us briefly consider them.

Firstly, there are privacy costs. Per Chapter 2, personalisation requires data. Various forms of personalisation, particularly those of the targeting and adaptive form, require access to increasingly intimate information about people, potentially to achieve outcomes that are not especially important (e.g., advertising). Indeed, various applications of AI in behavioural science suffer from this challenge, as alluded to in, say, the discussion of digital cloning (see Chapter 3). That chatbots may collect data which might infer latent behavioural traits, or that digital clones require personal data to be created (and may create personally consequential data), raise the costs of these applications in terms of privacy.

Secondly, there are the costs of the technological infrastructure needed for personalisation. This in-part relates to wider observations about the inefficiencies of means-tested policy provisions. It can be costly to identify who should receive a public provision, and who should not, and these costs may be more expensive than the money saved from being selective (Korpi and Palme, 1998). In the case of personalised behavioural science, AI systems and other algorithms are proposed as the means of predicting and targeting, with devices like smartphones providing the digital interface. This technological infrastructure does not come free, and this cost must be weighed against any benefit which comes from personalisation. Furthermore, the cost is increasingly not the economic cost of the hardware, but the environmental cost of the software, too. Modern AI models rely on huge data centres

[254] For instance, there is likely to be significant, meaningful variation in people's *abilities* to save, owing to income levels and spending needs.

[255] One might also argue, from a discrimination perspective, *not* personalising was preferable—is it right to draw exception because of someone's mental health?

which consume vast amounts of electricity and water, often to the determent of local communities (Hogan, 2015; Lehuedé, 2024; Monserrate, 2022). It would be a sad irony if AI systems were used to support behavioural science interventions to, say, reduce individual carbon emissions or take out home insurance, if that same infrastructure eroded the emissions benefits, or threatened citizens' houses with floods and wildfires.

Thirdly, there are the costs of explainability. Explainability has been discussed quite extensively in this chapter, but it is worth distilling the argument down and stating it clearly. Introducing an autonomous system into a decision-making process, or into the process of architecting choices, can obfuscate the chain of explainability and undermine citizen autonomy (Mills and Sætra, 2024). Personalisation is the example *par excellence*. An impersonal intervention is experienced by everyone. Some benefit more than others, but all experience the same intervention. This creates an opportunity for dialogue and social activity. People can discuss what happened. They can debate the fairness of it, or not. And then they can *reject* it (or celebrate it) as a group (Rawls, 1971). This 'publicity principle' was seen in Chapter 4 with the UK Government's grading algorithm. That the algorithm was applied for all effected children meant that the social inequities of the algorithm became a shared experience, facilitating the public debate about the acceptability of the approach. When people experience personalised interventions, there are limits to the common discourse one can undertake (Mills, 2022a). Rather than asking *is this good?* people must instead navigate the question *why did I get this, and you get that?* Only when (and if) people can answer *this* question (a question of facts) might the valuable discussion of values be debated.

Accepting these costs, in conjunction with the social and ethical concerns outlined in this chapter, it should not be regarded as evident that the use of autonomous systems in behavioural science is a means to desirable outcomes. These technologies might produce benefits, but there are also many reasons to be cautious.

Before the End

This chapter has dealt with autonomous systems, and specifically, autonomous choice architecture. It has tried to outline how the question of autonomy changes and frustrates behavioural science, raising important questions about the social and ethical implications of AI technologies when used to influence people's decisions. Combined with Chapter 4, this chapter has outlined the basis of normative objections to the use of AI in behavioural science; objections which are likely to be of growing importance to the field as technologies like autonomous choice architecture and adaptive nudging further develop.

Chapter [6]—Conclusion

> "It is terribly important to appreciate that some things remain obscure to the bitter end"
>
> —Stafford Beer, *Management Science* (1968, p. 115)

Choices

At the beginning of this book, I gave a definition of intelligent behaviour—*the process of selecting options predicted to achieve one's goals*. I then suggested to the reader that they might choose to *stop* reading at that point. In accordance with this definition, and assuming one's goal in reading this book was to learn about intelligent behaviour; having then and there being given a workable definition, continuing to read may not have been predicted to achieve one's subsequent goals. So, *did you stop reading*? Though, if one is reading *this*, it seems reasonable to assume that one has read this book beyond the opening page. As such, perhaps the intelligent questions to ask are *why did you keep reading?* and *was it an intelligent decision?*

All that remains is to conclude, which places a difficult decision at *my* door—having written the preceding chapters, what more should one say? Intelligent behaviour might suggest, as I see the hands on my clock pass firmly into the 'too late' territory, that I should write a short conclusion, summarising the main points while trying to finish on some seemingly profound line or call-to-action. Let us see what happens.

The question of intelligence, and intelligent behaviour, is one that is presently gripping the minds of people all over the world. As I write, I see that the Nobel Laureates of 2024 are debating AI and the role for humanity. The recipient for physics, Geoffrey Hinton, has proclaimed his profound concern of what might transpire when machines become more intelligent than people. The chemistry recipient, Demis Hassabis, has retorted that while AI systems continue to become 'smarter,' the challenge is ultimately one of human cooperation, not machine domination. The recipient for medicine, Gary Ruvkun, has then intervened, expressing his scepticism about human cooperation, and proclaiming people to be terribly overrated. Finally, the recipient for economics, Daron Acemoglu, has countered that in reality, people are tremendously *underrated* in modern society.

What is one to make of this discussion? For a start, one could reflect on the degree to which these distinguished minds actually *agree*—about the transformative and present dangers of AI technologies. The agreement reveals the fascinating component of their disagreement—over the role and the potential of people, not necessarily of machines. This book would raise some objections to some of the commentary around AI systems. For instance, to what extent is it helpful to talk of 'machines getting smarter than us' when, per Chapter 4, intelligence is a *social* product. This book would also express some sympathies. Most notably, around the complexities of people and the organisations in which people arrange themselves. It is interesting that in all the commentary on AI systems which one might encounter, the point of debate which makes these conversations fascinating, is *us*.

This book has set out, in part, to be a reference text for those in behavioural science and related fields. Chapters 2 and 3 have outlined example applications of AI technologies as they impact behavioural science. Chapters 4 and 5 have presented broad social and ethical arguments which intersect with AI and behavioural science, and which undoubtedly will crop up in one guise or another as this marriage develops. But an undercurrent of this book is that to truly grasp an application or an issue, one must constantly readjust one's focus, zoom in and out, change perspective, and other camera-based metaphors.

Intelligent behaviour can be understood individually and abstractly—as an individual choosing from a set of options to pursue a goal the origin of which is treated as exogenous. Though, it can also be understood socially (or interpersonally), as a game of influence between people, or as decisions taken by an individual whose goals and perceptions of options are a product of their socialisation in the world. Finally, one can consider the question organisationally (or institutionally), with intelligent behaviour being the product of people operating within a nexus of influence and culture, power and constraint. In each instance, how AI is *understood* may change, but what AI is (an information technology) remains the same. The complexity, and the intrigue, continues to be *us*.

As a reference book for behavioural science, though, this book may also fail. There is a degree of radicalism throughout this book (radical, by the standards of behavioural science) which has sought not merely to interrogate AI technology within intelligent behaviour, but *also* behavioural science itself. That people exhibit behavioural biases will sit comfortably with the behavioural scientist; that the notion of bias influences how behavioural scientists themselves understand people's behaviour, their motives and their goals, is a meta-commentary which sits less comfortably. This was a choice I have made, and I believe the book is better for it.

Decisions

It is a cliché to say that technology is neutral; that the goodness or badness of technology is what we choose to do with it. This does not stop it being more or less accurate, and it is accurate for AI technology, too. This book has presented many ways in which behavioural science can use AI technologies to understand and support people. Personalisation is a major application, and this is reflected in the substantial amount of space dedicated to the topic, notably in Chapter 2. As an information technology, information management is another substantial application within behavioural science, hence the focus of Chapter 3.

If one were to discuss the problems of AI technologies, or AI ethics, one could do so without any special reference to behavioural science. A myriad of AI-centric books furnish pages with references to philosophical thought experiments and the worse indulgences of science fiction. I hope this is not the impression given by this book. The decision to treat behavioural science as a social technology, and both AI and behavioural science as *political* creatures, has been done not merely to be contrarian or 'radical,' but because doing so allows one to consider novel problems which surround AI and behavioural science. Selective adherence, machine laundering, the assumption of error, the value-bias problem; these are some of the issues this book has focused on. Undoubtedly, there will be others. And undoubtedly, the analysis offered here will remain incomplete.

Though, as stated at the beginning, the question of intelligent behaviour is too broad to ever treat both comprehensively and meaningfully. I have attempted to exercise my own intelligent behaviour, and selected a narrower set of considerations which I predict will allow me to accomplish my goal. As I now conclude this book, I believe I have.

References

Acar, O. A., van den Ende, J. (2016) 'Knowledge Distance, Cognitive-Search Processes, and Creativity: The Making of Winning Solutions in Science Contests' *Psychological Science*, 27(5), pp. 692-699

Acemoglu, D., Autor, D., Born, D., Hanson, G. H., Price, B. (2014) 'Return of the Solow Paradox? IT, Productivity, and Employment in US Manufacturing' *The American Economic Review*, 104(5), pp. 394-399

Acemoglu, D., Johnson, S. (2023) '*Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*' Basic Books: UK

Agnew, W. A., Bergman, S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., McKee, K. R. (2024) 'The Illusion of Artificial Inclusion' *Proceedings of the CHI'24 Conference on Human Factors in Computing Systems*, 2024, pp. 1-12, DOI: 10.1145/3613904.3642703

Aher, G., Arriaga, R. I., Kalai, A. T. (2023) '*Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies'* Arxiv. Available at: https://arxiv.org/pdf/2208.10264.pdf

Alon-Barkat, S., Busuioc, M. (2023) 'Human-AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice' *Journal of Public Administration Research and Theory*, 33, pp. 153-169

Amirova, A., Fteropoulli, T., Ahmed, N., Cowie, M. R., Leibo, J. Z. (2023) '*Framework-Based Qualitative Analysis of Free Responses of Large Language Models: Algorithmic Fidelity'* Arxiv. Available at: https://arxiv.org/abs/2309.06364

Anderson, C. (2008) '*The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*' Wired. Available at: https://www.wired.com/2008/06/pb-theory/

Aoki, N. (2020) 'An experimental study of public trust in AI chatbots in the public sector' *Government Information Quarterly*, 37(4), e. 101490

Aonghusa, P. M., Michie, S. (2020) 'Artificial Intelligence and Behavioral Science Through the Looking Glass: Challenges for Real-World Application' *Annals of Behavioral Medicine*, 54, pp. 942-947

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., Wingate, D. (2023) 'Out of One, Many: Using Language Models to Simulate Human Samples' *Political Analysis*, 31(3), pp. 337-351

Arulsamy, K., Delaney, L. (2022) 'The impact of automatic enrolment on the mental health gap in pension participation: Evidence from the UK' *Journal of Health Economics*, 86, e. 102673

Ashby, W. R. (1978) '*Design for a Brain*' Chapman and Hall: UK

Attwell, K., Freeman, M. (2015) 'I Immunise: An evaluation of a values-based campaign to change attitudes and beliefs' *Vaccine*, 33(46), pp. 6235-6240

Bacon, L., Krpan, D. (2018) '(Not) Eating for the environment: The impact of restaurant menu design on vegetarian food choice' *Appetite*, 125(6), pp. 190-200

Barr, N., Pennycook, G., Stolz, J. A., Fugelsang, J. A. (2015) 'The brain in your pocket: Evidence that Smartphones are used to supplant thinking' *Computers in Human Behavior*, 48, pp. 437-480

Bashkirova, A., Krpan, D. (2024) 'Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance' *Computers in Human Behavior: Artificial Humans*, 2(1), e. 1000666

Bauer, J. M., Bietz, S., Rauber, J., Reisch, L. A. (2021) 'Nudging healthier food choices in a cafeteria setting: A sequential multi-intervention field study' *Appetite*, 160(5), e. 105106

Beer, S. (1968) '*Management Science: The Business Use of Operations Research*' Aldus Books: UK

Behavioural Insights Team (2023) '*AI chatbots in public services*' BIT. Available at: https://www.bi.team/wp-content/uploads/2024/01/External-AI-Chatbot-Trial-For-presentations.pdf

Bellamy, C. (2024) '*Experimenting with how generative AI could help GOV.UK users*' Inside Gov UK. Available at: https://insidegovuk.blog.gov.uk/2024/01/18/experimenting-with-how-generative-ai-could-help-gov-uk-users/

Ben-Aaron (1985) '*Weizenbaum examines computers and society*' The Tech. Available at: https://web.archive.org/web/20211002104454/http://tech.mit.edu/V105/N16/weisen.16n.html

Benartzi, S. (2017) '*The Smarter Screen: Surprising Ways to Influence and Improve Online Behavior*' Portfolio Profile: USA

Benartzi, S., Thaler, R. H. (1995) 'Myopic Loss Aversion and the Equity Premium Puzzle' *The Quarterly Journal of Economics*, 110(1), pp. 73092

Beshears, J., Kosowsky, H. (2021) 'Nudging: Progress to date and future directions' *Organizational Behavior and Human Decision Processes*, 161, pp. 3-19

Biss, E. (2022) '*The Theft of the Commons*' The New Yorker. Available at: https://www.newyorker.com/culture/essay/the-theft-of-the-commons

Blake, T., Nosko, C., Tadelis, S. (2015) 'Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment' *Econometrica*, 83(1), pp. 155-174

Blut, M., Wünderlich, N. V., Brock, C. (2024) 'Facilitating retail customers' use of AI-based virtual assistants: A meta-analysis' *Journal of Retailing*, 100(2), pp. 293-315

Blyth, M. (2013) 'Paradigms and Paradox: The Politics of Economic Ideas in Two Moments of Crisis' *Governance*, 26(2), pp. 197-215

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., Kalai, A. (2016) '*Man is to Computer Programmer as Women is to Homemaker? Debiasing Word Embeddings'* Available at: https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Bonabeau, E. (2002) 'Agent-based modelling: Methods and techniques for simulating human systems' *Proceedings of the National Academy of Science*, 99, pp. 7280-7287

Bostrom, N. (2014) '*Superintelligence: Paths, Dangers, Strategies*' Oxford University Press: UK

Bouschery, S. G., Blazevic, V., Piller, F. T. (2023) 'Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models' *Journal of Product Innovation Management*, 40, pp. 139-153

Brand, J., Israeli, A., Ngwe, D. (2023) '*Using GPT for Market Research'* Harvard Business School Marketing Unit Working Paper No. 23-062. SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4395751

Braverman, H. (1974) '*Labor and Monopoly Capitalism*' Monthly Review: USA

Brignull, H. (2011) '*Dark Patterns: Deception vs. Honesty in UI Design*' Available at: https://alistapart.com/article/dark-patterns-deception-vs.-honesty-in-ui-design/

Brontë, C. (2001) '*Jane Eyre*' W. W. Norton: UK

Brown, A. L., Grodzicki, D., Medina, P. C. (2022) 'When Nudges Spill Over: Student Loan Use under the CARD Act' SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4129413

Brunet, M., Alkalay-Houlihan, C., Anderson, A., Zemel, R. (2019) 'Understanding the Origins of Bias in Word Embeddings' *Proceedings of the 36th International Conference on Machine Learning*. DOI: 10.48550/arXiv.1810.03611

Bryan, C. J., Tipton, E., Yeager, D. S. (2021) 'Behavioural science is unlikely to change the world without a heterogeneity revolution' *Nature Human Behaviour*, 5, pp. 980-989

Bryne, C., Theakston, K., Randall, N. (2020) '*Gavin Williamson, Ofqual and the great A-level blame game'* The Conversation. Available at: https://theconversation.com/gavin-williamson-ofqual-and-the-great-a-level-blame-game-144766

Brynjolfsson, E., Collis, A., Liaqat, A., Kutzman, D., Garro, H., Deisenroth, D., Wernerfelt, N. (2024) '*The Consumer Welfare Effects of Online Ads: Evidence from a 9-Year Experiment*' NBER. Available at: https://www.nber.org/papers/w32846

Bucher, S., Dayan, P. (2023) '*Algorithmic Choice Architecture for Boundedly Rational Consumers*' Available at: https://static1.squarespace.com/static/61d37f2283265c45eac8ea1c/t/6547a44f7162ed27c4a4c5f8/1699193937496/Bucher_JMP.pdf

Buyalskaya, A., Ho, H., Milkman, K. L., Li, X., Duckworth, A. L., Camerer, C. (2023) 'What can machine learning teach us about habit formation? Evidence from exercise and hygiene' *Proceedings of the National Academy of Science*, 120(17), e. 2216115120

Buyalskaya, A., Gallo, M., Camerer, C. F. (2021) 'The golden age of social science' *Proceedings of the National Academy of Science*, 118(5), e. 2002923118

Caliskan, A., Bryson, J. J., Narayanan, A. (2017) 'Semantics derived automatically from language corpora contain human-like biases' *Science*, 356(6334), pp. 183-186

Camerer, C. F. (1989) 'Does the Basketball Market Believe in the 'Hot Hand'?' *The American Economic Review*, 79(5), pp. 1257-1261

Camerer, C. F. (1987) 'Do Biases in Probability Judgment Matter in Markets? Experimental Evidence' *The American Economic Review*, 77(5), pp. 981-997

Castelo, N., Boegershausen, J., Hildebrand, C., Henkel, A. P. (2023) 'Understanding and Improving Consumer Reactions to Service Bots' *Journal of Consumer Research*, 50(4), pp. 848-863

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., Banaji, M. R. (2021) 'Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words' *Psychological Science*, 32(2), pp. 218-240

Chater, N., Loewenstein, G. (2023) 'The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray' *Behavioral and Brain Sciences*, 46, e. 147

Chomsky, N. (1959) 'A Review of B. F. Skinner's Verbal Behavior' *Language*, 35(1), pp. 26-58

Chopra, F., Haaland, I. (2023) '*Conducting Qualitative Interviews with AI*' SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4572954

Clark, A., Chalmers, D. (1998) 'The Extended Mind' *Analysis*, 58(1), pp. 7-19

Cohen, M. D., March, J. G., Olsen, J. P. (1972) 'A Garbage Can Model of Organizational Choice' *Administrative Science Quarterly*, 17(1), pp. 1-25

Cummings, M. L. (2015) '*Automation Bias in Intelligent Time Critical Decision Support Systems*' in Harris, D. (eds.) '*Decision Making in Aviation*' (2015). Routledge: USA

Curchin, K. (2017) 'Using Behavioural Insights to Argue for a Stronger Social Safety Net: Beyond Libertarian Paternalism' *Journal of Social Policy*, 46(2), pp. 231-249

Cyert, R. M., March, J. G. (1992) '*A Behavioral Theory of the Firm*' Wiley: USA

Dalecke, S., Karlsen, R. (2020) 'Designing Dynamic and Personalized Nudges' *WIMS'20*. DOI: https://www.doi.org/10.1145/34059623405975

Daniels, G. S. (1952) '*The "Average Man"?*' Available at: https://www.dropbox.com/s/bsedsqvgbohy5wg/The%20%22Average%20Man%22%3F.pdf?dl=0

Darmody, A., Zwick, D. (2020) 'Manipulate to empower: Hyper-relevance and the contradictions of marketing in the age of surveillance capitalism' *Big Data and Society*, 7(1). DOI: https://doi.org/10.1177/2053951720904112

Davies, D. (2024) '*The Unaccountability Machine*' Profile: UK

De Cosmo, L. (2022) '*Google Engineer Claims AI Chatbot Is Sentient: Why That Matters*' Scientific American. Available at: https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/

Della Vigna, S., Linos, E. (2022) 'RCTs to Scale: Comprehensive Evidence From Two Nudge Units' *Econometrica*, 90(1), pp. 81-116

Dhami, S., Sunstein, C. R. (2023) '*Bounded Rationality: Heuristics, Judgment, and Public Policy*' MIT Press: USA

Dolan, P., Galizzi, M. M. (2015) 'Like ripples on a pond: Behavioral spillovers and their implications for research and policy' *Journal of Economic Psychology*, 47, pp. 1-16

Dolan, P., Hallsworth, M., Halpern, D., King, D., Metcalfe, R., Vlaev, I. (2012) 'Influencing behaviour: The mindspace way' *Journal of Economic Psychology*, 33(1), pp. 264-277

Drucker, P. F. (2006) '*The Effective Executive*' Harper Collins: USA

Duranton, G., Turner, M. A. (2011) 'The Fundamental Law of Road Congestion: Evidence from US Cities' *The American Economic Review*, 101(6), pp. 2616-2652

Edwards, L., Veale, M. (2017) 'Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For' *Duke Law and Technology Review*, 16, pp. 18-84

Egelman, S., Peer, E. (2015) 'The Myth of the Average User' *NSPW'15*, DOI: https://dx.doi.org/10.1145/2841113.2841115

Ellsberg, D. (2017) '*The Doomsday Machine: Confessions of a Nuclear War Planner*' Bloomsbury: UK

Evenepoel, A. (2022) '*Word Embeddings Quantify Stigmatising Attitudes and Stereotypes about Mental Illness*' Unpublished Manuscript.

Farahat, A., Bailey, M. C. (2012) 'How effective is targeted advertising?' *WWW '12: Proceedings of the 21st international conference on World Wide Web*, pp. 111-120. DOI: https://doi.org/1145.2187836.2187852

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., Netzer, O., Siegel, A. A., Plank, B., van Bavel, J. J. (2025) 'Using natural language processing to analyse text data in behavioural science' *Nature Reviews Psychology*. DOI: https://doi.org/10.1038/s44159-024-00392-z

Feyerabend, P. (1978) '*Science in a Free Society*' Verso Books: UK

Frederik, J., Martijn, M. (2019) '*The new dot com bubble is here: it's called online advertising*' The Correspondent. Available at: https://thecorrespondent.com/100/the-new-dot-com-bubble-is-here-its-called-online-advertising

Frey, C. B. (2019) '*The Technology Trap: Capital, Labor, and Power in the Age of Automation*' Princeton University Press: USA

Friedman, A. L. (1987) '*Industry and Labour: Class Struggle at Work and Monopoly Capitalism*' Macmillan: UK

Friestad, M., Wright, P. (1994) 'The Persuasion Knowledge Model: How People Cope with Persuasion Attempts' *Journal of Consumer Research*, 21(1), pp. 1-31

Frischmann, B., Selinger, E. (2018) '*Re-engineering Humanity*' Cambridge University Press: UK

Fuller, C. G. (2020) 'Uncertainty, insecurity, individual relative autonomy and the emancipatory potential of Galbraithian economics' *Cambridge Journal of Economics*, 44, pp. 229-246

Furr. M. R. (2009) 'Personality Psychology as a Truly Behavioural Science' *European Journal of Personality*, 23, pp. 369-401

Garcez, A. D., Lamb, L. C. (2020) '*Neurosymbolic AI: The 3rd Wave*' Available at: https://arxiv.org/abs/2012.05876v2

Garnelo, M., Shanahan, M. (2019) 'Reconciling deep learning with symbolic artificial intelligence: representing objects and relations' *Current Opinion in Behavioral Sciences*, 29, pp. 17-23

Geiger, G. (2021) '*How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud*' Vice. Available at: https://www.vice.com/en/article/how-a-discriminatory-algorithm-wrongly-accused-thousands-of-families-of-fraud/

Gerlich, M. (2025) 'AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking' *Societies*, 15(1). DOI: https://doi.org/10.3390/soc15010006

Gigerenzer, G. (2015) 'On the Supposed Evidence for Libertarian Paternalism' *Review of Philosophy and Psychology*, 6, pp. 361-383

Gigerenzer, G. (2007) '*Gut Feelings: Short Cuts to Better Decision Making*' Penguin Books: UK

Gitelman, L. (2013) '*Raw Data Is An Oxymoron*' MIT Press: USA

Glickman, M., Sharot, T. (2024) 'How human-AI feedback loops alter human perceptual, emotional and social judgements' *Nature Human Behaviour*. DOI: https://doi.org/10.1038/s41562-024-02077-2

Gmyrek, P., Lutz, C., Newlands, G. (2024) '*A Technological Construction of Society: Comparing GPT-4 and Human Respondents for Occupational Evaluation in the UK*' SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4700366

Goddard, K., Roudsari, A., Wyatt, J. C. (2011) 'Automation bias: a systematic review of frequency, effect mediators, and mitigators' *Journal of the American Medical Informatics Association*, 19, pp. 121-127

Golovianko, M., Gryshko, S., Terziyan, V., Tuunanen, T. (2023) 'Responsible cognitive digital clones as decision-makers: a design science research study' *European Journal of Information Systems*, 32(5), pp. 879-901

Golovianko, M., Gryshko, S., Terziyan, V., Tuunanen, T. (2021) 'Towards digital cognitive clones for the decision-makers: adversarial training experiments' *Procedia Computer Science*, 180, pp. 180-189

Graeber, D. (2015) '*The Utopia of Rules*' Melville: UK

Graeber, D. (2014) '*Debt: The First 5,000 Years*' Melville: UK

Gramsci, A. (2011) '*Prison Notebooks*' Columbia University Press: USA

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., Toombs, A. (2018) 'The dark (patterns) side of UX design' *CHI 2018*, 1-14

Greene, J. A., Lea, A. S. (2019) 'Digital Futures Past The Long Arc of Big Data in Medicine' *New England Journal of Medicine*, 381(5), pp. 480-485

Gregory, M., Tosi, A., McDonald, S., Sewell, R. (2024) '*The findings of our first generative AI experiment: GOV.UK Chat*' Inside Gov UK. Available at:

https://insidegovuk.blog.gov.uk/2024/01/18/the-findings-of-our-first-generative-ai-experiment-gov-uk-chat/

Gunkel, D. J. (2017) 'Mind the gap: responsible robotics and the problem of responsibility' *Ethics and Information Technology*, 22, pp. 307-320

Hadar, L., Sood, S. (2014) 'When Knowledge Is Demotivating: Subjective Knowledge and Choice Overload' *Psychological Science*, 25(9), pp. 1739-1747

Haenlein, M., Kaplan, A. (2019) 'A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence' *California Management Review*, 61(4), pp. 5-14

Hagar, N., Diakopoulos, N. (2019) 'Optimizing Content with A/B Headline Testing: Changing Newsroom Practices' *Media and Communication*, 7(1), pp. 117-127

Haig, B. D. (2014) '*Investigating the Psychological World: Scientific Method in the Behavioral Sciences*' MIT Press: USA

Haigh, T. (2023) '*There Was No 'First AI Winter*' ACM Communications. Available at: https://cacm.acm.org/opinion/there-was-no-first-ai-winter/

Hall, P. A. (1993) 'Policy paradigms, social learning, and the state: The case of economic policymaking in Britain' *Comparative Politics*, 25(3), pp. 275-296

Hallsworth, M. (2023) 'A manifesto for applying behavioural science' *Nature Human Behaviour*, 7, pp. 310-322

Hallsworth, M. (2022) '*Making Sence of the "Do Nudges Work?" Debate*' Behavioral Scientist. Available at: https://behavioralscientist.org/making-sense-of-the-do-nudges-work-debate/

Hämäläinen, P., Tavast, M., Kunnari, A. (2023) 'Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study' *CHI'23,* 2023, pp. 1-19. DOI: https://doi.org/10.1145/3544548.3580688

Harding, W., Kloster, M. (2024) '*Coding on Copilot: 2023 Data Shows Downward Pressure on Code Quality*' GitClear. Available at: https://www.gitclear.com/coding_on_copilot_data_shows_ais_downward_pressure_on_code_quality

Haselton, M. G., Nettle, D., Murray, D. R. (2015) '*The Evolution of Cognitive Bias*' in '*The Handbook of Evolutionary Psychology*' (2015), Wiley: UK

Hauser, J. R., Liberali, G., Urban, G. L. (2014) 'Website Morphing 2.0: Switching Costs, Partial Exposure, Random Exit, and When to Morph' *Management Science*, 60(6), pp. 1594-1616

Hauser, J. R., Urban, G. L., Liberali, G., Braun, M. (2009) 'Website Morphing' *Marketing Science*, 28(2), pp. 202-223

Hayek, F. A. (1999) '*The Sensory Order: An Inquiry into the Foundations of Theoretical Psychology*' University of Chicago Press: USA

Hayek, F. A. (1945) 'The Use of Knowledge in Society' *The American Economic Review*, 35(4), pp. 519-530

Hayes, A. F. (2017) '*Introduction to Mediation, Moderation, and Conditional Process Analysis*' Routledge: UK

Hecht, C. A., Dweck, C. S., Murphy, M. C., Kroeper, K. M., Yeager, D. S. (2023) 'Efficiently exploring the causal role of contextual moderators in behavioral science' *Proceedings of the National Academy of Science*, 120(1), e. 2216315120

Heidegger, M. (2010) '*Being and Time*' SUNY Press: USA

Heukelom, F. (2014) '*Behavioral Economics: A History*' Cambridge University Press: UK

Heukelom, F. (2012) 'Three Explanations for the Kahneman-Tversky Programme of the 1970s' *The European Journal of the History of Economic Thought*, 19(5), pp. 797-828

Hogan, M. (2015) 'Data flows and water woes: The Utah Data Center' *Big Data and Society*, 2(2), pp. 1-12

Illich, I. (1978) '*The Right to Useful Unemployment*' Marion Boyars: USA

Illich, I. (1973) '*Tools for Conviviality*' Marion Boyars: USA

Illich, I., Sanders, B. (1988) '*The Alphabetization of the Popular Mind*' Marion Boyars: UK

Ingrams, A., Kaufmann, W., Jacobs, D. (2021) 'In AI we trust? Citizen perceptions of AI in government decision making' *Policy and Internet*, 14, pp. 390-409

Iwasaki, M. (2024) 'Digital Cloning of the Dead: Exploring the Optimal Default Rule' *Asian Journal of Law and Economics*, 15(1), pp. 1-29

Jachimowicz, J. M., Duncan, S., Weber, E. U., Johnson, E. J. (2019) 'When and why defaults influence decisions: a meta-analysis of default effects' *Behavioural Public Policy*, 3(2), pp. 159-186

Jaffe, S., Shah, N. P., Butler, J., Farach, A., Cambon, A., Hecht, B., Schwarz, M., Teevan, J. (2024) '*Generative AI in Real-World Workplaces: The Second Microsoft Report on AI and Productivity Research*' Microsoft. Available at: https://www.microsoft.com/en-us/research/uploads/prod/2024/07/Generative-AI-in-Real-World-Workplaces.pdf

Jelveh, Z., Kogut, B., Naidu, S. (2024) 'Political Language in Economics' *The Economic Journal*, 134(8), pp. 2439-2469

Jensen, M. C., Meckling, W. H. (1976) 'Theory of the firm: Managerial behavior, agency costs and ownership structure' *Journal of Financial Economics*, 3(4), pp. 305-360

Johnson, E. J. (2021) '*How Netflix's Choice Engine Drives Its Business*' Behavioral Scientist. Available at: https://behavioralscientist.org/how-the-netflix-choice-engine-tries-to-maximize-happiness-per-dollar-spent_ux_ui/

Johnson, E. J., Goldstein, D. G. (2003) 'Do Defaults Saves Lives?' *Science*, 302, pp. 1338-1339

Joque, J. (2022) '*Revolutionary Mathematics: Artificial Intelligence, Statistics, and the Logic of Capitalism*' Verso Books: UK

Jussupow, E., Benbasat, I., Heinzl, A. (2020) 'Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion' *Proceedings of the 28th European Conference on Information Systems.* Available at: https://aisel.aisnet.org/ecis2020_rp/168/

Kahneman, D. (2011) '*Thinking, Fast and Slow*' Penguin Books: UK

Kahneman, D., Knetsch, J. L., Thaler, R. H. (1991) 'Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias' *The Journal of Economic Perspectives*, 5(1), pp. 193-206

Kahneman, D., Tversky, A. (1982) '*Intuitive prediction: Biases and corrective procedures*' in Kahneman, D., Slovic, P., Tversky, A. (eds.) '*Judgment under uncertainty: Heuristics and biases*' Cambridge University Press: UK

Kahneman, D., Tversky, A. (1979) 'Prospect Theory: An Analysis of Decision Under Risk' *Econometrica*, 47(2), pp. 263-291

Kaiser, M., Lohmann, P., Ochieng, P., Shi, B., Sunstein, C. R., Reisch, L. A. (2024) '*Leveraging LLMs for Predictive Insights in Food Policy and Behavioral Interventions*' Arxiv. Available at: https://arxiv.org/pdf/2411.08563

Katwala, A. (2020) '*An Algorithm Determined UK Students' Grades. Chaos Ensued*' Wired. Available at: https://www.wired.com/story/an-algorithm-determined-uk-students-grades-chaos-ensued/

Kaur, S., Mullainathan, S., Oh, S., Schilbach, F. (2024) 'Do Financial Concerns Make Workers Less Productive?' *The Quarterly Journal of Economics*. DOI: https://doi.org/10.1093/qje/qjae038

Keynes, J. M. (2017) '*The General Theory of Employment, Interest and Money*' Wordsworth: UK

Khambatta, P., Mariadassou, S., Morris, J., Wheeler, C. S. (2023) 'Tailoring recommendation algorithms to ideal preferences makes users better off' *Scientific Reports*, 13, e. 9325

Kim, J., Yoon, Y., Choi, J., Dong, H., Soman, D. (2024) 'Surprising Consequences of Innocuous Mobile Transaction Reminders of Credit Card Use' *Journal of Interactive Marketing*, 59(2), pp. 135-150

Kinchin, N., Mougouei, D. (2022) 'What Can Artificial Intelligence Do for Refugee Status Determination? A Proposal for Removing Subjective Fear' *International Journal of Refugee Law*, 34, 3-4, pp. 373-397

Kingdon, J. A. (2003) '*Agendas, Alternatives, and Public Policies*' Longman: USA

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S. (2018) 'Human Decisions and Machine Predictions' *The Quarterly Journal of Economics*, 133(1), pp. 237-293

Knetsch, J. L. (1989) 'The Endowment Effect and Evidence for Nonreversible Indifference Curves' *The American Economic Review*, 79(5), pp. 1277-1284

Kolkman, D. (2020) '*F**k the algorithm?: What the world can learn from the UK's A-level grading fiasco*' LSE Impact Blog. Available at: https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/

Kordzadeh, N., Ghasemaghaei, M. (2022) 'Algorithmic bias: review, synthesis, and future research directions' *European Journal of Information Systems*, 31(3), pp. 388-409

Korpi, W., Palme, J. (1998) 'The Paradox of Redistribution and Strategies of Equality: Welfare State Institutions, Inequality, and Poverty in the Western Countries' *American Sociological Review*, 63(5), pp. 661-687

Kosinski, M. (2021) 'Facial recognition technology can expose political orientation from naturalistic facial images' *Scientific Reportsi, 11,* 100.

Kosinski, M., Stillwell, T., Graepel, D. (2013) 'Private traits and attributes are predictable from digital records of human behavior' *Proceedings of the National Academy of Sciences*, 110(15), pp. 5802-5805

Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., Herzog, S. (2021) 'Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States' *Humanities and Social Sciences Communications*, 8, e. 117

Krefeld-Schwalb, A., Sugerman, E. R., Johnson, E. J. (2024) 'Exposing omitted moderators: Explaining why effect sizes differ in the social sciences' *Proceedings of the National Academy of Science*, 121(12), e. 2306281121

Kuhn, T. S. (2012) '*The Structure of Scientific Revolutions*' University of Chicago Press: USA

Lades, L. K., Delaney, L. (2022) 'Nudge FORGOOD' *Behavioural Public Policy*, 6(1), pp. 75-94

Laffan, K., Sunstein, C. R., Dolan, P. (2024) 'Facing it: assessing the immediate emotional impacts of calorie labelling using automatic facial coding' *Behavioural Public Policy*, 8(3), pp. 572-589

Lambert, C. (2015) '*Shadow Work: The Unpaid, Unseen Jobs That Fill Your Day*' Counterpoint: UK

Lanier, J. (2011) '*You Are Not A Gadget*' Penguin Books: UK

Lawrence, N. D. (2024) '*The Atomic Human: Understanding Ourselves in the Age of AI*' Allen Lane: UK

Ledley, R. S., Lusted, L. B. (1959) 'Reasoning Foundations of Medical Diagnosis: Symbolic logic, probability, and value theory aid our understanding of how physicians reason' *Science*, 130(3366), pp. 9-21

Lee, S., Peng, T., Goldberg, M. H., Rosenthal, S. A., Kotcher, J. E., Maibach, E. W., Leiserowitz, A. (2023) '*Can large language models capture public opinion about global warming? An empirical assessment of algorithmic fidelity and bias*' Arxiv. Available at: https://arxiv.org/ftp/arxiv/papers/2311/2311.00217.pdf

Lehuedé, S. (2024) 'An elemental ethics for artificial intelligence: water as resistance within AI's value chain' *AI and Society*, DOI: 10.1007/s00146-024-01922-2

Lepore, J. (2021) '*If Then: How One Data Company Invented the Future*' John Murray: UK

Levy, S. (2024) '*A Wave of AI Tools Is Set to Transform Work Meetings*' Wired Magazine. Available at: https://www.wired.com/story/taking-baby-steps-toward-the-ai-meeting-singularity/

Lewis, M. (2016) '*The Undoing Project: A Friendship that Changed the World*' Allen Lane: UK

Liberali, G., Ferecatu, A. (2022) 'Morphing for Consumer Dynamics: Bandits Meet Hidden Markov Models' *Marketing Science*, 41(4), pp. 663-869

Logg, J. M., Minson, J. A., Moore, D. A. (2019) 'Algorithm appreciation: People prefer algorithmic to human judgment' *Organizational Behavior and Human Decision Processes*, 151, pp. 90-103

Lohmann, P. M., Gsottbauer, E., Gravert, C., Reisch, L. A. (2024) '*Nudging fast and slow: Experimental evidence from food choices under time pressure*' CEBI Working Paper 19/24. Available at: https://www.econ.ku.dk/cebi/publikationer/working-papers/CEBI_WP_19-24.pdf

Longoni, C., Cian, L., Kyung, E. J. (2023) 'Algorithmic Transference: People Overgeneralize Failures of AI in the Government' *Journal of Marketing Research*, 60(1), pp. 170-188

Ludwig, J., Mullainathan, S. (2022) '*Algorithmic Behavioral Science: Machine Learning as a Tool for Scientific Discovery*' Working Paper no. 22-15. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4164272

Ludwig, J., Mullainathan, S. (2021) 'Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System' *Journal of Economic Perspectives*, 35(4), pp. 71-96

Luo, X., Rechardt, A., Sun, G., Nejad, K. K., Yañez, F., Yilmaz, B., Lee, K., Cohen, A. O., Borghesani, V., Pashkov, A., Marinazzo, D/. Nicholas, J., Salatiello, A., Sucholutsky, I., Minervini, P., Razavi, S., Rocca, R., Yusifov, E., Okalova, T., Gu, N., Ferianc, M., Khona, M., Patil, K. R., Lee, P., Mata, R., Myers, M. E., Bizley, J. K., Musslick, S., Bilgin, I. P., Niso, G., Ales, J. M., Gaebler, M., Murty, N. A. P., Loued-Khenissi, L., Behler, A., Hall, C. M., Dafflon, J., Bao, S. D., Love, B. C. (2024) 'Large language models surpass human experts in predicting neuroscience results' *Nature Human Behaviour*, DOI: https://doi.org/10.1038/s41562-024-02046-9

Lusted, L. B. (1960) 'Logical Analysis in Roentgen Diagnosis Memorial Fund Lecture' *Radiology*, 74(2), pp. 178-193

Lyell, D., Coiera, E. (2017) 'Automation bias and verification complexity: a systematic review' *Journal of the American Medical Informatics Association*, 24(2), pp. 423-431

Madrian, B. C., Shea, D. F. (2001) 'The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior' *The Quarterly Journal of Economics*, 116(4), pp. 1149-1187

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., Smolander, K. (2022) 'What influences algorithmic decision-making? A systematic literature review on algorithm aversion' *Technological Forecasting and Social Change*, 175, e. 121390

Maier, M., Bartoš, F., Stanley, T. D., Shanks, D. R., Harris, A. J. L., Wagenmakers, E. (2022) 'No evidence for nudging after adjusting for publication bias' *Proceedings of the National Academy of Science*, 119(31), e. 2200300119

Marcellino, W., Beauchamp-Mustafaga, N., Kerrigan, A., Chao, L. N., Smith, J. (2023) '*The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0*' RAND. Available at: https://www.rand.org/content/dam/rand/pubs/perspectives/PEA2600/PEA2679-1/RAND_PEA2679-1.pdf

March, J., Simon, H. A. (1993) '*Organizations*' Wiley: USA

Marcuse, H. (2013) '*One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society*' Beacon: USA

Marr, B. (2024) '*How Generative AI is Revolutionizing Customer Service*' Forbes. Available at: https://www.forbes.com/sites/bernardmarr/2024/01/26/how-generative-ai-is-revolutionizing-customer-service/

Martin, F. (2015) '*Money: The Unauthorised Biography*' Vintage: UK

Marx, K. (2013) '*Capital*' Wordsworth: UK

Matz, S. C., Kosinski, M., Nave, G., Stillwell, D. J. (2017) 'Psychological targeting as an effective approach to digital mass persuasion' *Proceedings of the National Academy of Science*, 114(48), pp. 12714-12719

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., Cerf, M. (2024) 'The potential of generative AI for personalized persuasion at scale' *Scientific Reports*, 14, e. 4692

McCarthy, J. (2007) '*What is Artificial Intelligence?*' Available at: http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

McCulloch, W. S. (1943) 'A logical calculus of the ideas immanent in nervous activity' *The Bulletin of Mathematical Biophysics*, 5, pp. 115-133

McKenzie, C. R. M., Liersch, M. J., Finkelstein, S. R. (2006) 'Recommendations Implicit in Policy Defaults' *Psychological Science*, 17(5), pp. 414-420

Medina, E. (2014) '*Cybernetic Revolutionaries: Technology and Politics in Allende's Chile*' MIT Press: USA

Mei, Q., Xie, Y., Yuari, W., Jackson, M. O. (2024) 'A Turing test of whether AI chatbots are behaviorally similar to humans' *Proceedings of the National Academy of Science*, 121(9), e. 2313925121

Meijer, A., Lorenz, L., Wessels, M. (2021) 'Algorithmization of Bureaucratic Organizations: Using a Practice Lens to Study How Context Shapes Predictive Policing Systems' *Public Administration Review*, 81(5), pp. 837-846

Merchant, B. (2023) '*Blood in the Machine: The Origins of the Rebellion Against Big Tech*' Little and Brown: UK

Moynihan, T. (2020) '*X-Risk: How Humanity Discovered Its Own Extinction*' Urbanomic: UK

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013) '*Efficient Estimation of Word Representations in Vector Space*' Arxiv. Available at: https://arxiv.org/abs/1301.3781

Miller, G. A. (2003) 'The Cognitive Revolution: A Historical Perspective' *Trends in Cognitive Science*, 7(3), pp. 141-144

Miller, G. A. (1956) 'The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information' *Psychological Review*, 101(2), pp. 343-352

Mills, S. (2024a) 'Being Good and Doing Good in Behavioral Policymaking' *Public Administration Review*. DOI: https://doi.org/10.1002/puar.13908

Mills, S. (2024b) '*Algorithms, Bytes, and Chips: The Emerging Political Economy of Foundation Models*' SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4834417

Mills, S. (2022a) 'Personalized Nudging' *Behavioural Public Policy*, 6(1), pp. 150-159

Mills, S. (2022b) 'Finding the 'nudge' in hypernudge' *Technology in Society*, 71, e. 102117

Mills, S., Costa, S., Sunstein, C. R. (2023) 'AI, Behavioural Science, and Consumer Welfare' *Journal of Consumer Policy*, 46, pp. 387-400

Mills, S., Sætra, H. S. (2025) '*Algorithms in the Room: AI, Representation, and Decisions about Sustainable Futures*' Unpublished Manuscript.

Mills, S., Sætra, H. S. (2024) 'The Autonomous Choice Architect' *AI and Society*, 39, pp. 583-595

Mills, S., Spencer, D. A. (2025) 'Efficient Inefficiency: Organisational challenges of realising economic gains from AI' *Journal of Business Research*, 189, e. 115128

Mills, S., Whittle, R. (2025) '*How Nudge Happened: The Political Economy of Behavioural Insights in the UK*' *Cambridge Journal of Economics*, 49(1), pp. 1-18

Mills, S., Whittle, R. (2024a) 'Seeing the nudge from the trees: The 4S framework for evaluating nudges' *Public Administration*, 102(2), pp. 580-600

Mills, S., Whittle, R. (2023) '*Detecting Dark Patterns Using Generative AI: Some Preliminary Results*' SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4614907

Minsky, M. L. (1986) '*The Society of Mind*' Picador: USA

Minsky, M. L., Papert, S. A. (2017) '*Perceptrons: An Introduction to Computational Geometry*' MIT Press: USA

Mittelstadt, B., Russell, C., Wachter, S. (2019) 'Explaining Explanations in AI' *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 19, pp. 279-288

Monserrate, S. G. (2022) '*The Staggering Ecological Impacts of Computation and the Cloud*' The MIT Press Reader. Available at: https://thereader.mitpress.mit.edu/the-staggering-ecological-impacts-of-computation-and-the-cloud/

Moon, Y. (2010) '*Different: Escaping the Competitive Herd*' Crown Currency: USA

Moon, Y. (2002) 'Personalization and Personality: Some Effects of Customizing Message Style Based on Consumer Personality' *Journal of Consumer Psychology*, 12(4), pp. 313-325

Morozov, E. (2013) '*To Save Everything, Click Here: Technology, Solutionism, and the Urge to Fix Problems that Don't Exist*' Allen Lane: UK

Morozovaite, V. (2022) 'Hypernudging in the changing European regulatory landscape for digital markets' *Policy and Internet*, 15(1), pp. 78-99

Morozovaite, V. (2021) 'Two Sides of the Digital Advertising Coin: Putting Hypernudging into Perspective' *Market and Competition Law Review*, 5(2), pp. 105-145

Mosier, K., Dunbar, M., McDonnell, L., Skitka, L., Burdick, M., Rosenblatt, B. (1998) 'Automation bias and errors: Are teams better than individuals?' *Proceedings of the Human Factors and Erogonomics Society Annual Meeting*, 42(3), pp. 201-205

Mrkva, K., Posner, N. A., Reeck, C., Johnson, E. J. (2021) 'Do Nudges Reduce Disparities? Choice Architecture Compensates for Low Consumer Knowledge' *Journal of Marketing*, 85(4), pp. 67-84

Mullainathan, S., Obermeyer, Z. (2022) 'Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care' *The Quarterly Journal of Economics*, 137(2), pp. 679-727

Mullainathan, S., Obermeyer, Z. (2017) 'Does Machine Learning Automate Moral Hazard and Error?' *American Economic Review: Papers and Proceedings 2017*, 107(5), pp. 476-480

Murakami, K., Shimada, H., Ushifusa, Y., Ida, T. (2022) 'Heterogeneous treatment effects of nudge and rebate: Causal machine learning in a field experiment on electricity conservation' *Internation Economic Review*, 63(4), pp. 1779-1803

Murphy, M. (2024) 'Artificial Intelligence and Personality: Large Language Models' Ability to Predict Personality Type' *Emerging Media*, 2(2), pp. 311-324

Muthukrishna, M. (2023) '*A Theory of Everyone*' Basic Books: UK

Muthukrishna, M., Henrich, J. (2016) 'Innovation in the collective brain' *Philosophical Transactions of the Royal Society B*, 371, e. 20150192

Narayanan, S., Yu, G., Ho, C., Yin, M. (2023) 'How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making?' *AAAI/ACM Conference on AI, Ethics, and Society*. DOI: https://doi.org/10.1145/3600211.3604709

Nazaretsky, T., Cukurova, M., Ariely, M, Alexandron, G. (2021) '*Confirming bias and trust: Human factors that influence teachers' attitudes towards AI-based educational technology*' Available at: https://discovery.ucl.ac.uk/id/eprint/10141423/

Negroponte, N. (1995) '*Being Digital*' Knopf: USA

Newall, P. W. S. (2019) 'Dark nudges in gambling' *Addiction Research and Theory*, 27(2), pp. 65-67

Newell, A., Simon, H. A. (2019) '*Human Problem Solving*' Echo Point: USA

Nisa, C. F., Sasin, E. M., Faller, D. G., Schumpe, B. M., Bélanger, J. J. (2020) 'Reply to: Alternative meta-analysis of behavioural interventions to promote action on climate change yields different conclusions' *Nature Communications*, 11(3901), pp. 1-3

Nye, P., Thomson, D. (2020) '*A-Level results in 2020: Why independent schools have done well out of this year's awarding process*' Education Data Lab. Available at: https://ffteducationdatalab.org.uk/2020/08/a-level-results-2020-why-independent-schools-have-done-well-out-of-this-years-awarding-process/

Obermeyer, Z., Lee, T. H. (2017) 'Lost in thought: The limits of the human mind and the future of medicine' *New England Journal of Medicine*, 377(13), pp. 1209-1211

Ofqual (2024) '*A level outcomes in England*' Available at: https://analytics.ofqual.gov.uk/apps/Alevel/Outcomes/

Ofqual (2020) '*Written statement from Chair of Ofqual to the Education Select Committee*' UK Government. Available at: https://www.gov.uk/government/news/written-statement-from-chair-of-ofqual-to-the-education-select-committee

O'Toole, J. (2024) '*Generative Design: Shaping a More Adaptive Web*' Available at: https://www.cbmdigital.co.uk/blog/generative-design-shaping-a-more-adaptive-web

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., Bernstein, M. S. (2024) '*Generative Agent Simulations of 1,000 People*' Arxiv. Available at: https://arxiv.org/pdf/2411.10109

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., Bernstein, M. S. (2023) 'Generative Agents: Interactive Simulacra of Human Behavior' *ACM UIST'23*. DOI: https://doi.org/10.1145/3586183.3606763

Park, J. S., Popowski, L., Cai, C. J., Morris, M. R., Liang, P., Bernstein, M. S. (2022) '*Social Simulacra: Creating Populated Prototypes for Social Computing Systems*' Arxiv. Available at: https://arxiv.org/pdf/2208.04024

Pasquinelli, M. (2023) '*The Eye of the Master: A Social History of Artificial Intelligence*' Verso Books: UK

Pearson, J. (2020) '*The Story of How the Australian Government Screwed Its Most Vulnerable People*' Vice. Available at: https://www.vice.com/en/article/the-story-of-how-the-australian-government-screwed-its-most-vulnerable-people-v27n3/

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' *Computers in Human Behavior*, 109, e. 106347

Peer, E., Mills, S. (2024) '*From One, Many: A Framework for Personalised Nudging*' Available at: https://osf.io/b7fd9

Peters, B. (2017) '*How Not to Network a Nation: The Uneasy History of the Soviet Internet*' MIT Press: USA

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J. H., Mertz, C. K. (2007) 'Less Is More in Presenting Quality Information to Consumers' *Medical Care Research and Review*, 64(2), pp. 169-190

Peters, E., Markowitz, D. M. (2024) '*Numbers Are Persuasive—If Used in Moderation*' Scientific American. Available at: https://www.scientificamerican.com/article/numbers-are-persuasive-if-used-in-moderation/

Peters, E., Markowitz, D. M., Nadratowski, A., Shoots-Reinhard, B. (2024) 'Numeric social-media posts engage people with climate science' *PNAS Nexus*, 3(7), e. 250

Peterson, A. J. (2024) '*AI and the Problem of Knowledge Collapse*' Arxiv. Available at: https://arxiv.org/abs/2404.03502

Petracca, E. (2021) 'On the origins and consequences of Simon's modular approach to bounded rationality in economics' *The European Journal of the History of Economic Thought*, 28(5), pp. 708-732

Pettifor, A. (2019) '*The Case for the Green New Deal*' Verso Books: UK

Polanyi, K. (2024) '*The Great Transformation: The Political and Economic Origins of Our Time'* Penguin Books: UK

Polanyi, M. (2005) '*Personal Knowledge: Towards a Post-Critical Philosophy*' Routledge: UK

Possati, L. M. (2020) 'Algorithmic unconscious: why psychoanalysis helps in understanding AI' *Palgrave Communications*, 6, p. 70

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., Barnes, P. (2020) '*Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing'* Arxiv. Available at: https://arxiv.org/abs/2001.00973

Rauthmann, J. F. (2020) 'A (More) Behavioural Science of Personality in the Age of Multi-Modal Sensing, Big Data, Machine Learning, and Artificial Intelligence' *European Journal of Personality*, 34, pp. 593-598

Rawls, J. (1971) '*A Theory of Justice*' Harvard University Press: USA

Reinecke, K., Gajos, K. Z. (2014) 'Quantifying Visual Preferences Around the World' *CHI'14*. DOI: https/dx.doi/10.1145/2556288.2557052

Reñosa, M. D. C., Landicho, J., Wachinger, J., Dalglish, S. L., Bärnighausen, K., Bärnighausen, T., McMahon, S. A. (2021) 'Nudging toward vaccination: a systematic review' *BMJ Global Health*, 6, e. 006237

Rheault, L., Cochrane, C. (2020) 'Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora' *Political Analysis,* 28(1), pp. 112-133

Ridgway, V. F. (1956) 'Dysfunctional Consequences of Performance Measurements' *Administrative Science Quarterly*, 1(2), pp. 240-247

Risko, E. F., Gilbert, S. J. (20166) 'Cognitive Offloading' *Trends in Cognitive Sciences*, 20(9), pp. 676-688

Rosenblatt, F. (1962) '*Principles of Neurodynamics*' Spartan Books: USA

Russell, S. J. (2023) 'AI weapons: Russia's war in Ukraine shows why the world must enact a ban' *Nature*, 614, pp. 620-623

Russell, S. J. (2019) '*Human Compatible: AI and the Problem of Control*' Penguin Books: UK

Russell, S. J. (1997) 'Rationality and Intelligence' *Artificial Intelligence*, 94, pp. 57-77

Russell, S. J., Norvig, P. (2020) '*Artificial Intelligence: A Modern Approach*' Prentice-Hall: USA

Sætra, H. S. (2023) 'Generative AI: Here to stay, but for good?' *Technology in Society*, 75, e. 102372

Sætra, H. S. (2019) 'When nudge comes to shove: Liberty and nudging in the era of big data' *Technology in Society*, 59, e. 101130

Sætra, H. S., Selinger, E. (2023) '*The Siren Song of Technological Remedies for Social Problems: Defining, Demarcating, and Evaluating Techno-Fixes and Techno-Solutionism*' SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4576687

Salvaggio, E. (2023) '*Flowers Blooming Backward Into Noise*' Available at: https://www.cyberneticforests.com/news/flowers-blooming-backward-into-noise-2023

Samoili, S., López-Cobo, M., Gómez, E., de Prato, G., Martínez-Plumed, F., Delipetrev, B. (2020) '*AI Watch Defining Artificial Intelligence: Towards an operational definition and taxonomy of artificial intelligence*' JRC Technical Reports EUR 30117 EN. Available at: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC118163/jrc118163_ai_watch._defining_artificial_intelligence_1.pdf

Samuelson, W., Zeckhauser, R. (1988) 'Status Quo Bias in Decision Making' *Journal of Risk and Uncertainty*, 1, pp. 7-59

Sanders, M., Snijders, V., Hallsworth, M. (2018) 'Behavioural science and policy: where are we now and where are we going?' *Behavioural Public Policy*, 2(2), pp. 144-167

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T. (2023) 'Whose Opinions Do Language Models Reflect?' *Proceedings of the 40th International Conference on Machine Learning.*

Scheibehenne, B., Greifeneder, R., Todd, P. M. (2010) 'Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload' *Journal of Consumer Research*, 37(3), pp. 409-425

Schelling, T. C. (1980) '*The Strategy of Conflict*' Harvard University Press: USA

Schultz, P. W., Messina, A., Tronu, G., Limas, E. F., Gupta, R., Estrada, M. (2016) 'Personalized Normative Feedback and the Moderating Role of Personal Norms: A Field Experiment to Reduce Residential Water Consumption' *Environment and Behavior*, 48(5), pp. 686-710

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N.J., Griskevicius, V. (2007) 'The Constructive, Destructive, and Reconstructive Power of Social Norms' *Psychological Science*, 18(5), pp. 429-434

Scott, L. (2015) '*The Four-Dimensional Human: Ways of Being in the Digital World*' Penguin Books: UK

Scriven, M. (1953) 'The Mechanical Concept of Mind' *Mind*, 62(246), pp. 230-240

Selten, F., Robeer, M., Grimmelikhuijsen, S. (2023) ''Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment' *Public Administration Review*, 83(2), pp. 263-278

Sen, A. (2002) '*Rationality and Freedom*' Harvard Belknapp: USA

Sher, S., McKenzie, C. R. M. (2006) 'Information leakage from logically equivalent frames' *Cognition*, 101(3), pp. 467-494

Sher, S., McKenzie, C. R. M., Müller-Trede, J., Leong, L. (2022) 'Rational Choice in Context' *Current Directions in Psychological Science*, 31(6), pp. 518-525

Shrestha, P., Krpan, D., Koaik, F., Schnider, R., Sayess, D., Binbaz, M. S. (2024) '*Beyond WEIRD: Unveiling GPT's Cultural Bias in Global Policy Research'* Unpublished Manuscript.

Skidelsky, R. (2023) '*The Machine Age*' Allen Lane: UK

Skidelsky, R., Skidelsky, E. (2012) '*How Much Is Enough? The Love of Money, and the Case for the Good Life*' Allen Lane: UK

Skinner, B. F. (1984) 'An operant analysis of problem solving' *Behavioral and Brain Sciences*, 7, pp. 583-613

Skinner, B. F. (1976) '*About Behaviorism*' Vintage: USA

Skitka, L. J., Mosier, K. L., Burdick, M. (2000) 'Accountability and automation bias' *International Journal of Human-Computer Studies*, 52(4), pp. 701-717

Skitka, L. J., Mosier, K. L., Burdick, M. (1999) 'Does automation bias decision-making?' *International Journal of Human-Computer Studies*, 51(5), pp. 991-1006

Silver, D., Singh, S., Precup, D., Sutton, R. S. (2021) 'Reward is enough' *Artificial Intelligence*, 209, e. 103535

Simon, H. A. (1997a) '*Administrative Behavior*' Free Press: USA

Simon, H. A. (1997b) '*An Empirically Based Microeconomics*' Cambridge University Press: UK

Simon, H. A. (1996) '*Models of My Life*' MIT Press: USA

Simon, H. A. (1987a) 'Two Heads Are Better than One: The Collaboration between AI and OR' *Interfaces*, 17(4), pp. 8-15

Simon, H. A. (1987b) 'Making Management Decisions: The Role of Intuition and Emotion' *The Academy of Management Executive*, 1(1), pp. 57-64

Simon, H. A. (1981) '*The Sciences of the Artificial*' MIT Press: USA

Simon, H. A. (1956) 'Rational Choice and the Structure of the Environment' *Psychological Review*, 63(2), pp. 129-138

Simon, H. A. (1955) 'A Behavioral Model of Rational Choice' *The Quarterly Journal of Economics*, 69(1), pp. 99-118

Smith, A., Harvey, J., Goulding, J., Smith, Sparks, L. (2021) 'Exogenous cognition and cognitive state theory: The plexus of consumer analytics and decision-making' *Marketing Theory*, 21(1), pp. 53-74

Snow, T. (2021) 'From satisficing to artificing: The evolution of administrative decision-making in the age of the algorithm' *Data and Policy*, 3, e. 3

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., Choi, Y. (2024) '*A Roadmap to Pluralistic Alignment'* Arxiv. Available at: https://arxiv.org/abs/2402.05070

Sparrow, B., Liu, J., Wegner, D. M. (2011) 'Googe Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips' *Science*, 333, pp. 776-779

Špecián, P. (2023) '*Large Language Models for Democracy: Limits and Possibilities*' Technology and Sustainable Development 2023 Conference Paper. Available at: https://techandsd.com/_files/specian_2023.pdf

Sposini, F. M. (2019) 'At the borders of the average man: Adolphe Quêtelet on mental, moral and criminal monstrosities' *Journal of Historical Behavioural Science*, 1, pp. 1-17

Stadler, M., Bannert, M., Sailer, M. (2024) 'Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry' *Computers in Human Behavior*, 160, e. 108286

Sternberg, R. J. (1999) 'The Theory of Successful Intelligence' *Review of General Psychology*, 3(4), pp. 292-316

Stevenson, M. (2018) 'Assessing Risk Assessment in Action' *Minnesota Law Review*, 103, pp. 303-384

Stiglitz, J. (2024) '*The Road to Freedom*' Allen Lane: UK

Sunstein, C. R. (2024) 'Choice engines and paternalistic AI' *Humanities and Social Science Communications*, 11, e. 888

Sunstein, C. R. (2023) 'The use of algorithms in society' *Review of Austrian Economics*, DOI: https://doi.org/10.1007/s11138-023-00625-z

Sunstein, C. R. (2022a) 'The distributional effects of nudges' *Nature Human Behaviour*, 6, pp. 9-10

Sunstein, C. R. (2022b) 'Governing by Algorithm? No Noise and (Potentially) Less Bias' *Duke Law Journal*, 71(6), pp. 1175-2105

Sunstein, C. R. (2019) 'Algorithms, Correcting Biases' *Social Research: An International Quarterly*, 86(2), pp. 499-511

Sunstein, C. R. (2017) '*Misconceptions about Nudges*' SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3033101

Sunstein, C. R. (2015) 'The Ethics of Nudging' *Yale Journal of Regulation*, 32, pp. 413-450

Sunstein, C. R. (2014) 'Choosing Not to Choose' *Duke Law Journal*, 64, pp. 1-52

Sunstein, C. R. (2013a) '*Why Nudge? The Politics of Libertarian Paternalism*' Yale University Press: USA

Sunstein, C. R. (2013b) 'The Storrs Lectures: Behavioral Economics and Paternalism' *Yale Law Journal*, 122, pp. 1826-1899

Tegmark, M. (2017) '*Life 3.0: Being Human in the Age of Artificial Intelligence*' Penguin Books: UK

Thakur, S. (2024) '*Put Google AI to work with Search ads*' Google. Available at: https://blog.google/products/ads-commerce/put-google-ai-to-work-with-search-ads/

Thaler, R. H. (2021) 'What's next for nudging and choice architecture?' *Organizational Behavior and Human Decision Processes*, 163, pp. 4-5

Thaler, R. H. (2018) 'Nudge, not sludge' *Science*, 361(6401), pp. 431-432

Thaler, R. H. (2015) '*Misbehaving: The Making of Behavioural Economics*' Penguin Books: UK

Thaler, R. H., Sunstein, C. R. (2008) '*Nudge: Improving Decisions about Health, Wealth and Happiness*' Penguin Books: UK

Thaler, R. H., Sunstein, C. R. (2003) 'Libertarian Paternalism' *The American Economic Review*, 93(2), pp. 175-179

Thaler, R. H., Tucker, W. (2013) 'Smarter Information, Smarter Consumers' *Harvard Business Review*, 91(1-2), pp. 44-45

Thunström, L. (2019) 'Welfare effects of nudges: The emotional tax of calorie menu labelling' *Judgment and Decision Making*, 14(1), pp. 11-25

Thunström, L., Gilbert, B., Jones-Ritten, C. (2018) 'Nudges that hurt those already hurting – distributional and unintended effects of salience nudges' *Journal of Economic Behavior and Organization*, 153, pp. 267-282

Truby, J., Brown, R. (2021) 'Human digital thought clones: the *Holy Grail* of artificial intelligence for big data' *Information and Communications Technology Law*, 30(2), pp. 140-168

Turkle, S. (2013) '*Alone Together: Why We Expect More From Technology and Less From Each Other*' Basic Books: UK

Turkle, S. (2004) '*The Second Self: Computers and the Human Spirit*' MIT Press: USA

Turkle, S. (1988) ‘Artificial Intelligence and Psychoanalysis: A New Alliance’ *Dædalus*, 117(1), pp. 241-268

Tversky, A., Kahneman, D. (1992) ‘Advances in Prospect Theory: Cumulative Representation of Uncertainty’ *Journal of Risk Uncertainty*, 5, pp. 297-323

Tversky, A., Kahneman, D. (1983) ‘Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment’ *Psychological Review*, 90(4), pp. 293-315

Tversky, A., Kahneman, D. (1974) ‘Judgment under Uncertainty: Heuristics and Biases’ *Science*, 185(4156), pp. 1124-1131

Tversky, A., Kahneman, D. (1973) ‘Availability: A Heuristic for Judging Frequency and Probability’ *Cognitive Psychology*, 5, pp. 207-232

UK Behavioural Scientists (2020) ‘*Open letter to the UK Government regarding COVID-19*’ Available at: https://sites.google.com/view/covidopenletter/home

UK Government (2023) ‘*Generative AI framework for HM Government*’ HM Government. Available at: https://assets.publishing.service.gov.uk/media/65c3b5d628a4a00012d2ba5c/6.8558_CO_Generative_AI_Framework_Report_v7_WEB.pdf

Valcea, S., Hamdani, M. R. (2024) ‘Exploring the Impact of ChatGPT on Business School Education: Prospects, Boundaries, and Paradoxes’ *Journal of Management Education*, 48(5), pp. 915-947

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’ *Nature Communications*, 11(4553), pp. 1-12

Von Neumann, J. (2000) ‘*The Computer and The Brain*’ Yale University Press: USA

Von Neumann, J., Morgenstern, O. (1944) ‘*Theory of Games and Economic Behavior*’ Princeton University Press: USA

Wachter, S., Mittelstadt, B., Floridi, l. (2017) ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’ *International Data Privacy Law*, 7(2), pp. 76-99

Wang, Y., Kosinski, M. (2018) ‘Deep neural networks are more accurate than humans at detecting sexual orientation from facial images’ *Journal of Personality and Social Psychology*, 114(2), pp. 246-257

Weale, S., Adams, R. (2020) ‘*Durham University offers students money to defer entry*’ The Guardian. Available at: https://www.theguardian.com/education/2020/aug/19/durham-university-offers-students-money-to-defer-entry

Weizenbaum, J. (1976) ‘*Computer Power and Human Reason: From Judgment to Calculation*’ W. H. Freeman: USA

Wenzelburger, G., König, P. D., Felfeli, J., Achtziger, A. (2024) ‘Algorithms in the public sector. Why context matters’ *Public Administration*, 102(1), pp. 4-60

West, R., Michie, S., Chadwick, P., Atkins, L., Lorencatto, F. (2020) '*Achieving behaviour change: A guide for national government*' Public Health England. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/933328/UFG_National_Guide_v04.00__1___1_.pdf

Wiener, N. (1950) '*The Human Use of Human Beings*' Houghton Mifflin: USA

Wilke, A., Mata, R. (2012) '*Cognitive Bias'* in Ramachandran, V. S. (eds.) '*Encyclopaedia of Human Behaviour*' 2nd ed. (2012). Academic Press: USA

Wright, P. (2002) 'Marketplace Metacognition and Social Intelligence' *Journal of Consumer Research*, 28(4), pp. 677-682

Yeung, K. (2017) 'Hypernudge: Big Data as a mode of regulation by design' *Information, Communication and Society*' 20(1), pp. 118-136

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

Zhang, Y., Gosline, R. (2024) 'Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human-GAI collaboration in persuasive content generation' *Judgment and Decision Making*, 18, e. 41

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

Zuboff, S. (2015) 'Big other: surveillance capitalism and the prospects of an information civilization' *Journal of Information Technology*, 30, pp. 75-89