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.[1] ELIZA was a 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

[1] 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.

"small and simple step" in the development of natural language processing, *if that* (Weizenbaum, 1976, p. 7).[2]

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.[3] 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 of programming knowledge.[4] For many people, text generation of the sophistication of GPT-3 *was* novel, and was worth sharing.[5]

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 physical way in 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.[6] 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.

---

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

[3] 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.

[4] 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.'

[5] 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).

[6] 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*.

*

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 overvaluing the status quo (Samuelson and Zeckhauser, 1988) to relying too much on similarity, salience, and those ideas most speedily brought to mind (Tversky and Kahneman, 1974).[7] 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.[8]

Such a proposal is contentious. One controversy may be that the notion that algorithms *counteract* 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 deals with c-biases. Insofar as one focuses on c-biases, it is a responsible 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).[9]

Much has been said on the topic of algorithmic bias (e.g., Kordzadeh and Ghasemaghaei, 2022), and the claim that algorithms de-bias decision-makers is far from the only controversy which might arise from the widespread deployment of AI to support biased decision-makers. This chapter will begin with these other complaints, of which there are two.

---

[7] People are also regularly shown 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).

[8] Writing on the use of algorithms in medical settings, Greene and Lea (2019) offer a similar argument to that laid out 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 if *inherent* human biases.

[9] 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 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.

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 normative 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 challenging 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*.[10] 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 concerns that of 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.[11] 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,[12] 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

[10] I cannot claim that this is an original term, and one I have definitely stolen from someone else. Unfortunately, I cannot remember who I have stolen it from. To this end, no creative credit be given to me for this term.

[11] For instance, Dhami and Sunstein (2023) note that all cognitive biases in the Kahneman-Tversky research programme as essentially measurements of deviations from economic rationality. Of course, having a benchmark does not necessarily imply some normative judgement. At the same time, 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.

[12] 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.

would imagine a compelling majority in opposition to the practice of using a defendant's mugshot within a bail decision.[13]

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.[14]

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 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 a couple hundred of 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 risk of death upon their community, if bailed.

In the negative, though, one might launch the following argument: 1) the criminal justice system is designed to punish, reform, and 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.[15]

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

---

[13] 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.

[14] 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?*

[15] 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).

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 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.[16] 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.[17] Secondly, the alternative that did exist—the use of predicted grades given to each student by their teacher—was likely to not produce accurate outcomes.[18] 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 likely to be a poor predictor of actual student performance.[19] 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%.

---

[16] 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 to the bias it was designed to solve. The toeslagenaffaire and robodebt will be discussed, in a different context, later in this chapter.

[17] 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.

[18] 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.

[19] 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."

While the majority of grades—nearly 59%—remained unchanged from those predicted by teachers,[20] 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).[21] Private schools 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. 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.[22] 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.[23] And so on. Given inflated grades have important knock-on effects—for instance, placing greater strain on more competitive higher education institutions[24]—a behavioural scientist may support the use of an algorithmic model or

[20] 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)?

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

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

[23] 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).

[24] 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).

predictive AI tool to assign grades, at least in this special instance where students could not feasibly complete the exams.[25]

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

What this example demonstrates is the limits of behavioural science as a means of advocating for algorithms or AI in society. 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 therein expressed. 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

[25] All of this does raise 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 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, 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.

[26] 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 here is the exercise.

eroded.[27] 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

The question of why an algorithm or AI system might be used in decision-making is an expansive one. Let us expand on it further.

An interesting perspective emerges when one considers whether algorithms and AI technologies actually debias people. Snow (2021) interviews several policymakers about how algorithms are used within policymaking. From these interviews, Snow (2021) outlines what he calls *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.[28] On the other end, there is algorithm aversion, or the tendency to ignore algorithms even when algorithms lead to superior outcomes.[29] These are both interesting behaviours, with automation bias being the subject of some discussion in Chapter

[27] 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).

[28] 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). However, 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."

[29] 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.

5. However, the immediate discussion will focus on those behaviours which fall within the middle of the artificing spectrum.[30]

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. Alon-Barkat and Busuioc 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 oversight which the algorithm lacks.

However, the study reports 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 lead to 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 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 discretion is not used in a way which

[30] 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 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.

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) ask participants to evaluate case studies of patients seeking a new kidney and are tasked with making a recommendation to either approve a transplant, or to deny it—an ethically difficult, and socially important, decision. Participants are equipped with an AI algorithm to support their decision-making. Some participants receive an algorithm that has been given a set of ethical preferences which differ from the participants; others receive one which is 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 medical benefits *versus* the medical costs of doing so, 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. Nevertheless, 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. 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 would 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.[31]

Finally, Bashkirova and Krpan (2024) investigate the use of AI triage recommendations by psychologists working in the domain of mental healthcare. As with previous studies, Bashkirova and Krpan find that the psychologists tended 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

[31] 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.

expert. Bashkirova and Krpan note that perceived expertise may frustrate the introduction of AI algorithms in professional decision-making domains.[32]

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 in a biased way*?[33] 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.[34] 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 invent *post hoc* rationalisations for their decisions when challenged, because their behaviour is motivated by personal interests.[35]

A potentially useful framework to explore motivations for using, and overruling, algorithms can be found in Mills and Sætra's (2024b) 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.[36] 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 where failure comes from, 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 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

[32] 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.

[33] 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 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.

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

[35] 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.

[36] 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.

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.

They 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.[37] 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.[38] 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 (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, 2000). 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*.[39] 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.

[37] The reader is reminded of the idea of *exogenous cognition* mentioned in a prior footnote and encouraged to draw links between the above argument and various discussion found in Chapter 3 around information management.

[38] 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.

[39] For a thematically suitable study of how these factors influence policy adoption, see Mills and Whittle's (2024b) study of the political-economic factors which influenced the rise of nudging.

These factors which lead decision-makers to ignore perspectives are not *cognitive* biases. Instead, they are political and institutional factors.[40] 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 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 diverse 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.[41] Mills and Sætra describe such uses of AI technologies as *technosolutionism*, a term for coined by Morozov (2013) 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.

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.

A remarkably similar scandal, the *robodebt* scandal, unfolded in Australia in 2016. Here, an automated system erroneously claimed thousands of benefits claimants had been 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

[40] 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.

[41] 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."

as political scapegoats.[42] 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.[43]

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.

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 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—but only because it allows policymakers to avoid alternative solutions which they *already* do not want to pursue.[44]

---

[42] 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.

[43] 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).

[44] 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 (2024b) elaborate on the politics of nudging and behavioural insights in an attempt to demonstrate more completely the 'crowding out' perspective.

One can make similar machine laundering arguments about the other instances already discussed, involving doctors and teachers. For instance, a grading bias may actually be demonstrative of a lack of teaching resources (e.g., time) to accurately assess a student's prospects. Yet, given long-term funding constraints on schools, fiscal uncertainty from the pandemic and fiscal austerity in the years preceding it, providing these resources may have been politically undesirable, while individualising the problem (as a bias) and solving *that* problem with an algorithm became a solution.[45] In the process, what decision-makers wanted (or did *not* want) to do was laundered by an algorithm—transformed from a desired objective into an *objective* assessment of reality.

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.[46] 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.

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 itself*. The idea of machine laundering presages this point—that the motivation for adopting AI technology in decision-making might itself be based on reasoning which leads to inequitable outcomes. This section explores this phenomenon through the link between behavioural science and technology adoption within organisations. Doing so will also add helpful additional perspectives to the arguments already developed in this chapter.

There is a famous saying within the world of management, only ever half remembered given it is so obvious substantial attention seems to be a poor investment. It states that *what gets measured gets managed*. It is often attributed to Peter Drucker, the famed management consultant, and is often

---

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.

[45] 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*.

[46] 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.

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*.[47] Yet, Drucker never said *what gets measured gets managed*, 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.[48] Thus, drawing on Ridgway, 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 measurement,[49] he was sceptical that the same applied for knowledge work,[50] 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.[51] 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).[52]

---

[47] 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.

[48] 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).

[49] 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."

[50] 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.

[51] 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 contribution of 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.

[52] 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.

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 (2000) distinction between *value* judgements and *factual* judgements.[53] For Drucker (2006, p. 145), to ignore the role of opinions (or, in Simon's language, values) is to make a fatal mistake when important decisions must be taken.[54] 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."[55]

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 problem,[56] and that this solution came from a failure to truly examine how

---

[53] Simon (2000, 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 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."

[54] 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."

[55] 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.

[56] 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

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, an apparent technological solution is often preferred to an actual behavioural one—in this instance, increasing the number of cables printed (Simon, 2000).

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-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 the report. 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 *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

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).

offered, seized upon, and justified—in the mind of worker, and the organisation—by pointing to positive changes in whatever is measured.[57]

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; 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. 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. 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.

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, 2000). Individuals in organisations are themselves boundedly rational, and this character then echoes throughout the whole organisation (Cyert and March, 1963). 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, 1963). 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 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 need knowledge of how

[57] 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 perform 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.

processes and technologies work, hindering the discovery of ways in which they *could work better* (Acemoglu and Johnson, 2023).[58]

One worthwhile example is self-service kiosks. Self-service kiosks allow supermarkets to dismiss checkout operators. But these technologies do not replace the operator with a faster, more efficient machine. Instead, they merely *shift* who does 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 technological disintermediates the programmer, leaving the everyday user to stumble around to figure out what the technology can and cannot do, what it should and should not do, and so on. In such an environment, 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.[59]

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, accidentally and intentionally. 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. One should conclude, though, that

---

[58] 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.

[59] 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 itself, 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 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.

whether an algorithm *ought* to be used depends on much more than what an algorithm or AI system *can* do, or what a person *cannot* do.

Ultimately, the use of algorithms and AI systems in society is what we, as a society, make of them. Chapters 2 and 3 stressed a difference between the advocacy for algorithms and AI given by Simon (1987a, 1987b) and that given by Sunstein (2024, 2023, 2022b, 2019).[60] But, technically speaking, one might argue that there is little separating them. Simon emphasises information filtering and using algorithms as part of the design of an organisational solution to allow experts to prosper. Sunstein emphasises the biases of experts and using algorithms to offer correctives to experts given they are biased.

The difference, though, is how each implicitly frames the problem they are trying to solve. Simon sees the problem as being a problem of *organisation* and embraces all the complexities which this chapter has outlined. In Simon's work, the use of algorithms, AI, and computers in general should be part of a broad programme of organisational design.[61] Sunstein sees the problem as being a problem of the *individual* and may be accused of sidelining the same complexities through a loose advocacy for discretion. Advocacy is vital for the reasons he provides—because the world is uncertain and requires adaption which an algorithm may be ill-equipped to demonstrate. But much like the teleprinter, algorithms which focus on tackling biases risk doing little but diminishing the agency of the individual even further, frustrating rather than supporting the organisational advocacy which might be key to achieving superior outcomes.

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 was to roleplay as a therapist, responding to people to elicit 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."[62]

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 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

[60] Also see Mills (2024) and Mills and Spencer (2025).

[61] Hence Simon's (1981) call for a *science of design* or even the title of his book on the matter, *the sciences of the artificial*; the sciences of that which humankind makes.

[62] 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.

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."[63]

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.[64] 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.[65] The same is true of predictive AI, though here it is more subtle. Given enough

---

[63] 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.

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.

[64] 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*.

[65] 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."

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 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."[66] 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.[67] 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.[68]

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) then 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.

[66] 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.

[67] 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*."

[68] 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. Unsurprisingly, this idea has links to Marxist philosophy, particularly the notion of alienation—a detachment from the sociality of one's work and effort. See Braverman (1974).

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 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." It is interesting to consider that modern AI systems essentially assume the relation between reality, R, and data, D, is given by some equation $R = \gamma D$, where γ 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

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. Behavioural scientists must thus be careful in pushing this narrative.

As this chapter has sought to demonstrate, there are ample paths of human-AI collaboration along which the behavioural scientist could tread. This chapter has sought to offer a map.

$V = \gamma P$, whereby the market is said to serve the same function as the AI in estimating γ. Knowing γ, in both instances, supposes one does not need to know reality, or value, only data, and prices.