Chapter [3]—Machines Like Us

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

The Social Life of Information

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

That decisions can be impeded by too much information, and that information can be obscured, are not controversial findings within behavioural science. Yet, the manner in which information is discussed within behavioural science often diminishes what one might call the 'social life' of information and encourages one to view information in a way more analogous to computer science—as *bits* of information taking discrete values (e.g., one, or zero). Of course, we all know that information in the ordinary, everyday sense (and therefore, in the behavioural science sense) has a qualitative dimension to it, even if our language often emphasises the quantitative aspect.[1] This chapter will benefit from a brief discussion of these qualitative aspects of everyday information, or, as above, information's 'social life.'

Firstly, many will sympathise with the idea that sometimes one can have 'too much' information. In economics and psychology, *choice overload* can arise from decisions involving too much choice, and thus too many elements for a decision to consider (Hadar and Sood, 2014).[2] Moon (2010) has somewhat amusingly used choice overload and 'too much' information to explain why supermarkets often sell dozens of different brands of bottled water. Moon argues that by presenting too much choice, consumers become overwhelmed and incapable of making *any* choice, leaving them vulnerable to savvy salespeople who 'help' consumers by 'lending' them their 'expert' knowledge.[3] 'Too much' information also links to Miller's (1956) famous 'magic number seven'

[1] One might be tempted to construct some mental model linking information to data, data to knowledge, knowledge to wisdom, and so on. Such models are cute, but do not contribute a great deal to the ideas contained within this chapter. More interesting categories almost certainly emerges when one considers knowledge that can be recorded as information, and knowledge which cannot be (e.g., tacit knowledge; Polanyi, 1974).

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

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

paper, which suggests the average human can remember—and effectively manipulate—seven pieces ('bits' or 'chunks') of information in their short-term memory, plus or minus two.

While there is certainly a quantitative aspect to 'too much' information, all information *counts*. For instance, some choices will be readily eliminated from the choice set. It is not that people struggle to navigate many choices *per se*, but that those choices which one already struggles to distinguish may be harder to distinguish when there are many more of them.[4] The inequality of information puts an important spin on Miller's magic number seven, also. As Simon (1981) notes, the three-letter string QUV may require three 'chunks' of memory to remember, but CAT is likely to only require one, as is ONE and, ironically, TWO. These latter strings are the same as common words we already know, and thus can draw on associations to remember these words, rather than short-term 'chunks.'[5] QUV, by contrast, lacks those associations, requiring us to remember, separately, 'Q', 'U', and 'V'.[6] This belies another issue—sometimes it is not that information *must* be considered, but that information distracts one from what *ought to be* considered (Simon, 2000).

Second is the matter of obscurity. Often desirable information is available, but not easily. This can encourage one to seek out *more* information as an unhelpful attempt to remedy the problem (Simon, 1981). Such is often the case in business and government (Drucker, 2006). This problem is tied to that of too much information insofar as too much information can obscure and distract from relevant information, as immediately above. But the lack of easy availability of information is a wider category of impediment which undermines effective decision-making. For instance, the state can feasibly gain information on essentially any topic or domain. But some projects will face substantial budget costs and difficulties from citizens (perhaps because of budget costs) which will disincentivise the undertaking of the project, and the collection of relevant information (Kingdon, 2003). There is also the matter of *ambiguity*, which is taken to mean a scenario which can be explained by two wholly contradictory sets of facts. In ambiguous situations, a lack of availability obscures what information is available (Ellsberg, 2017).[7]

---

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

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

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

[5] In Simon's (1981) language, one could say that CAT is *isomorphic* to the word cat. Isomorphism is an important idea within Simon's brand of information processing and symbol manipulation. Two entities are isomorphic when the rules by which they operate are the same, but the *representations* of the entities are different. By recognising the similarities between entities, one need only remember the similarity, not the representation. This reduces the amount of information one needs to remember and manipulate.

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

[7] Ellsberg (2017) has defined ambiguity this way in his discussion of nuclear war planning. Say a nuclear bomber pilot is ordered to drill a nuclear attack plan. Wanting on accurate test, the pilot's commanders do not tell the pilots involved

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

This chapter will present examples of both. A reader is encouraged not to linger on this brief tangent into philosophy but is encouraged to think about this tangent in relation to the subsections which follow.

Filtering Information and Finding Biases

Perhaps the most important application of AI within behavioural science is as a data analysis tool. While this function is much less titillating than other applications, it hardly an outlandish proposition, and thus one which must be taken seriously. AI, and specifically predictive AI, presents novel ways of analysing data, and encourages the uses of different (and more) data. In doing so, predictive AI may support behavioural scientists both in their investigations of human behaviour and in their deployment of these insights to problems in everyday life.

Aonghusa and Michie (2020) discuss a fascinating application of AI in their work as part of the *Human Behaviour Change Project*. An enormous amount of literature exists examining various interventions to change habits and encourage healthier lifestyles. The amount of literature alone might be daunting, but the variety adds an additional dimension which creates informational challenges.[8] Different studies use different samples and sample sizes. They target different behaviours via different behavioural mechanisms by deploying different behavioural interventions. Some are experimental studies, others quasi-experimental, and so on. The heterogeneity of the literature frustrates one's abilities to draw coherent insights from the literature. It also undermines adaptiveness to real-world policy pressures, from unfolding public health crises to political changes which reorientate priors, budgets, and philosophies on public service.

The programme of work discussed by Aonghusa and Michie (2020) involves using an AI system to synthesise these many thousands of public health studies. Then, through a user interface, researchers can query the system for predictions based on the literature. As queries change, in response to the changing policy environment, the system can make new predictions, which policymakers can use to update their plans. Both arriving at a synthesised prediction of a public

---

that it is a drill. The diligent pilot loads up their nuclear payload, taxis along the runway, and proceeds towards the enemy city. The pilot knows that if this were a test, they would soon receive disarm orders from the base. Then, suddenly, there is a flash of light, followed by an enormous explosion. A nuclear blast on the base that the pilot has just taken off from. Ellsberg invites us to consider what the pilot should do. On the one hand, the pilot has been told nuclear war has begun, has been ordered to attack the enemy, and has just seen a nuclear explosion. *It all makes sense.* On the other hand, the explosion might have been caused by an accident involved as part of the drill. The blast has now killed the commanders who would have issued the disarm orders. War has not been declared, and the worst thing the pilot can do is to keep going. *It all makes sense.* Ellsberg points to the inherent ambiguity of this situation: two sensible but contradictory interpretations of the same information. The pilot's ignorance of what caused the blast obscures how they should interpret the information that is readily available to them.

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

health intervention and updating the prediction as circumstances around the intervention change, would be tremendously difficult for even a trained team of researchers to accomplish. Aonghusa and Michie (2020) note that the use of AI for these functions empowers public health experts to better utilise valuable research *which is available*, but which is not *readily accessible*. In a comparable and recent study, Kaiser *et al.* (2024) report that a fine-tuned large language model (LLM) can demonstrate a high predictive accuracy of behavioural experiments to change eating habits. This leads these researchers to suggest that, through further refinement and technological development, AI could become a vital information management and prediction tool for behavioural scientists and policymakers. Luo *et al.* (2024) report similar results from LLMs trained to predict neuroscience results.

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

Another area of AI application is likely to be in detecting behavioural biases (Mills, Costa and Sunstein, 2023). While this remains an area dominated more by speculation than practical results, there are robust arguments and *some* results which suggest this is a feasible application to consider.[9] Mills, Costa and Sunstein (2023) argue that one may understand a behavioural bias very much like a tendency within data. Consider the default bias. One might imagine that a dataset showing two choices, a default option (coded as '1') and an alternative option (coded as '0'), approximately half of observations would be registered as a '1' and half as a '0'. A significant deviation from this pattern, favouring '1' over '0', would be indicative of a bias towards the default, assuming the default option changes arbitrarily for different people, and assuming several instances of this experiment (with different choices, participants, and environments) demonstrate the same result. From this perspective, behavioural biases will arise as *patterns* in datasets. Predictive AI technologies, by design, spot and act upon patterns. Thus, the argument is that once behavioural biases are understood as patterns in data, AI technologies may help behavioural scientists identify known biases in novel settings, and perhaps more interestingly, new biases in old and new settings alike.

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

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

between words can be analysed mathematically. A common analysis is to examine the semantic similarity between different words.[10]

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

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

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

[10] *Word2Vec* was the most prominent of a group of word vectorisation applications developed in the mid-2010s. Ideas found within these models likely contributed in part to the development of today's large language models. It may be dubious to describe *Word2Vec* and similar as AI *per se*. Developed by Google engineers, *Word2Vec* essentially predicted patterns between words using a relatively basic neural network consisting of a single layer of neurons. Through training, this layer comes to 'embed' numerical information which captures the relationship between different words (without the numbers themselves necessarily meaning anything). This allows individual words to be represented as vectors of numbers (hence *vectorisation*), which can be probed mathematically to infer semantic relationships. For instance, semantic similarity (how similar are two words in everyday language) can be measured by observing the 'distance' between the vectors of the respective words in an *n*-dimension vector space, where *n* is the length of the vector.

This latter argument has been extended into the development of 'word embedding association tests' (WEATs), where word vectorisation is used to investigate implicit biases in groups, before being compared with implicit biases identified through established behavioural science methods, most notably the implicit association test (IAT). Caliskan *et al.* (2017) develop the WEAT approach and demonstrate that it can accurately replicate the implicit biases found in a separate IAT. In another study, Charlesworth *et al.* (2021) use word embeddings to demonstrate how gender stereotypes persist in big text data, showing how associations between gender and toys, media, occupations, and so on, follow expected patterns of stereotyping. Evenepoel (2022) extends the WEAT method further. By analysing big text datasets from different decades, Evenepoel (2022) shows how biases and attitudes have changed over time. For instance, one WEAT of attitudes towards depression demonstrates a declining association between depression and psychological causes, perhaps reflecting greater appreciation of the various causes of mental health, compared to popularly held beliefs in the mid-twentieth century.[11] It has even been shown that word embeddings can accurately predict the ideological positions of politicians by identifying common word associations found within different parts of the political spectrum (Rheault and Cochrane, 2020).

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

Alternate to word embedding approaches, others have used AI methods to identify a variety of data points which might subsequently be called biases. Various studies have investigated how doctors and judges make decisions around, say, deferring potential heart attack patients for further testing, or determining whether a defendant should be offered the opportunity for bail, respectively. By training predictive AI models on various data available to these decision-makers, researchers can then probe these models to determine which data points have an outsized predictive power within the model. Assuming these especially powerful variables do not conform to what one might consider 'rational' standards for decision-making, the predictive power of the variable may be indicative of biased behaviour on the part of the decision-maker (Ludwig and Mullainathan, 2022, 2021).[12]

For instance, evidence suggests that judges frequently exhibit two biases when making bail decisions, which Sunstein (2023) dubs 'current offence bias' and 'mugshot bias.' As these names imply, studies suggest that a defendant's current offence has an outsized role in whether they receive bail, when—one might argue—their *whole* offending history is a better indication of whether one is likely to commit another crime if granted bail (Kleinberg *et al.*, 2018). In the case of the mugshot, a defendant's mugshot has outsized predictive power when predicting if bail will be granted (Ludwig and Mullainathan, 2022), when a person's appearance should not have any bearing on their likelihood to commit a crime in the future.[13] Sunstein (2023) argues that both have

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

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

[13] These findings relating to social justice can be contentious. For instance, one may suggest that the current offence bias should not be considered as such. Assuming the criminal justice system serves its function of rehabilitating, and

links to the more 'fundamental' bias of availability bias (Tversky and Kahneman, 1974)—the tendency to focus on immediately available information (like a mugshot, or current offence) rather than less apparent information (like a defendant's past offending history).

A similar such conclusion can be reached when considering Mullainathan and Obermeyer's (2022) AI-led investigation of doctor referral procedures. They find that doctors tend to over-test low risk patients (sending them for a battery of tests which would determine severe illnesses they are unlikely to have) while under-testing high risk patients (who should be subjected to many tests which, compared to the average, might seem excessive). This, Mullainathan and Obermeyer report, leads to inefficiencies in healthcare provision, with many patients subject to waste (in terms of time and medical costs), and some subject to risk from undiagnosed or underdiagnosed symptoms. They also note that rational precautions—say, because doctors do not want to be sued—do not explain the pattern of under- and over-testing. Instead, doctors seem to evaluate a patient's health risk based on immediately available information, rather than a more complete medical history and circumstantial details which would inform a more accurate assessment of risk. As with Judges, this finding is the result of AI-based analysis techniques which allow influential decision-making factors to be investigated and analysed.[14]

Such insights can inform policy recommendations. For instance, judges could receive algorithmic recommendations as to whether a defendant is likely to reoffend, while doctors could be told of a patient's risk, as predicted by an AI model. The formatting of the information judges receive might also be reconfigured so that the most relevant information (in terms of best guidance) is also the most available information, while irrelevant information (like the mugshot) is less available, or even removed entirely (Ludwig and Mullainathan, 2021; Sunstein, 2023). So too might medical information. One worthwhile criticism of such recommendations, though, is the potential for AI recommendations to gain an outsized influence in decision-making processes which should be undertaken by a person.[15] To this end, one might return again to Simon's (1987a,

---

that once a punishment has been served a citizen is no longer bound to atone for their crime, then one could ask if it is not unfair to consider a defendant's whole past history? What is called a bias, in this instance, is merely a matter of perspective about the function of the criminal justice system.

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

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

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

1987b) advocacy for human-AI collaboration and AI as an 'expert support system.'[16] Here, insofar as AI can identify and remove distracting information which undermines the effective decision-making of a judge or a doctor, without the AI *actually* making the decision, may allow a happy medium to be reached where a human decides, but decides more consistently and equitably.[17]

Who Are You Talking To?

Online choice architecture (OCA) is a behavioural scientist's way of talking about a website, or what a user interface (UI) designer would call a *user interface* (obviously). This is not to suggest that OCA represents a proliferation of terms for its own sake, as if the behavioural scientists cannot play nicely with the UI designers. The utility of a term like OCA is it encourages one to see interfaces that are interacted with every day as drivers of individual choice rather than merely neutral receptacles of options. Much like 'analogue' behavioural science did not necessarily propose radically different ideas to those found in marketing, but did emphasise a different viewpoint from which insights could be gleamed, so too with 'digital' behavioural science and OCA.

The OCA space is growing quite considerably. To an extent, the intersection of AI with OCA is driven more by the pre-existing interest in technology found amongst OCA researchers, than by compelling evidence of the role of AI. Though, this is not a universal statement. Recommendation algorithms are a common component of OCA, typically found in a 'recommended for you' box on retail websites.[18] Unpicking designs and highlighting the behavioural science contained within them can also be time consuming and subject to human bias. Automating such 'auditing' processes may also be a domain where AI could supercharge OCA

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

As early as 1960, Lusted (1960) argued that computers would function as information filtering systems in medicine, splitting suspect patient scans off from non-suspect ones: "an electronic scanner-computer [will] look at chest photofluorograms, to separate the clearly normal chest films from the abnormal chest films. The abnormal chest films would be marked for later study by the radiologists." This is not dissimilar, in principle, to Simon's proposals—that AI should assist experts to do their jobs better by managing unhelpful information. According to Greene and Lea (2019, p. 480), Zworykin had "warned that medical data were accumulating at a pace exceeding physicians' cognitive capacity" as early as 1964. This followed perhaps one of the first information revolutions anywhere, with new technologies for diagnosis and information storage leading to rapid increases in medical data collection throughout the 1950s.

Interestingly, much of the concern about 'judgement' was not directed at *doctors'* judgement, but at the judgement of *programmers* of computers. This is not wholly absent from discussions today (e.g., Mullainathan and Obermeyer, 2017). Yet, recent behavioural science studies have emphasised the fallibility of *doctors* and the benefits of AI algorithms as a result. This is perhaps a meaningful inversion of what is—to reiterate—hardly a *new* area of discussion. I have argued that this inversion may be understood by examining the relatively 'narrow' view of bounded rationality popular in behavioural science today, compared to Simon's (2000, 1981) initial discussions of the idea (Mills, 2024).

To be sure, there was recognition of doctors' bias in the sense that doctors might disagree, preferring information which aligned to their specialty (Ledley and Lusted, 1959), though this is hardly a deviation from Simon's (2000) observations in the 1940s about information preferences in organisational decision-making. This initial emphasis on biases and disagreement remained orientated towards the challenge of *programming* a useful AI for doctors. Such observations were not necessarily used as *motivating* the use of AI technologies, as one might argue is the case today. As Greene and Lea (2019) note, Zworykin was concerned about such a conclusion and felt it necessary to emphasise that the goal was not to undermine doctors. Neither, do I think, is the objective of recent behavioural science studies. An emphasis on *doctors'* biases, nevertheless, risks doing so.

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

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

research and support regulatory goals of protecting customers and fostering greater competition amongst firms (Mills and Whittle, 2024). This is an application I will return to in the discussion of social simulations with AI. But perhaps the most interesting intersection involves chatbots powered by generative AI. This is a nascent area—one which may slip more into speculation—but of importance given the growing experimentation with generative AI and given existing OCA which perhaps makes AI sales assistants useful.

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

The AI chatbot can be understood as serving a similar purpose, at least in principle, while extending the capabilities through which a person can interact with a website to include natural language communication. Through natural language interactions with users of a website, an AI chatbot could serve as a value information management tool. A person could ask the chatbot to recommend products on a retail website, or to provide a link to an online form for claiming some government provision. The UK Government (2023), for instance, has outlined in its guidance to civil servants that the use of large language models (LLMs) as part of supporting citizen queries could be an appropriate application of AI. Its Government Data Service (GDS) has argued that "there is potential for [AI chatbots] to have a major, and positive, impact on how people use [government websites] – for instance making it easier to find answers to their questions from the 700,000+ page estate" (Bellamy, 2024, para. 3). A subsequent experiment run by the GDS found around two-thirds of users to be satisfied with an early AI chatbot on government webpages (Gregory *et al.*, 2024).[19]

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

Firstly, those given access to an AI chatbot performed worse on a multiple-choice task than a control group with no AI chatbot.[20] At the same time, only around 40% of those would access to the chatbot actually used it. One explanation of the poor performance may therefore be that for most people, the chatbot window was merely an additional distraction adding to the informational complexity to be navigated. This is perhaps demonstrated through a second finding.

Secondly, of those who *did* use the chatbot, accuracy did in fact improve in some instances.[21] Interestingly, those who used the chatbot were also *slower* in completing the multiple-choice task—

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

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

[21] The Behavioural Insights Team (2023) tested four different AI chatbot designs. The most intrusive designs (whole page chatbots) lead to worse performance or identical performance to the control group who had no access to an AI

though by only around three seconds. Experientially, this and the use of an AI chatbot in general appears to have been relatively minimal. Participants did not really find government information any more or less difficult to understand, though perceptions of task ease were higher in three of the four treatment groups, compared to the control.

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

While retailers may share some of the challenges and opportunities of AI chatbots with governments, they also face considerations unique to the private sector. One recent study (Castelo *et al.*, 2023) has found that consumers dislike AI chatbots compared to human customer service attendants, in part because consumers believe that private firms only use AI for their own private benefit, such as being able to cut down on staffing costs. The same study has found, though, that consumer attitudes improve when the AI chatbot produces a clear benefit for consumers, too, either through lower prices (e.g., discounts that only the chatbot can offer) or superior information (e.g., a significantly improved customer experience). This being so, there is hope that adoption of AI chatbots may improve—though it is questionable to what extent AI, rather than other improvements in online services, will encourage this adoption.[22]

Such findings belie an important aspects of AI chatbots which should return attention to older ideas about sales and persuasion. An AI chatbot placed on a retailer's website is analogous to a salesperson in a store. While both *may* be helpful (they are both typically called 'assistants,' after all), both the AI chatbot and the salesperson work for the retailer and are employed precisely because it is believed they increase sales.[23] Though, it is not always clear *how* this is achieved. Maybe both improve the customer experience, which indirectly increases sales by, say, encouraging customers to return more frequently than they would if they *just* liked the product? Or maybe both effectively deploy persuasion techniques to encourage the customer to purchase a product that they otherwise might not?

Such exercises in speculation are what marketers and consumer behaviour researchers call *metacognition*—thinking about thinking (Friestad and Wright, 1994; Wright, 2002). Metacognition is an activity we all engage in, and a frequent example is when one walks into a store and must have

chatbot. The less intrusive designs saw improvements over the control, though it is not clear if these improvements were statistically significant.

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

[23] At the time of writing, there does not appear to be clear empirical evidence that AI chatbots have a causal relationship with higher sales. Though, there is substantial discussion of AI chatbots within some industrial magazines and amongst various consultancies. For the purposes of discussion, I will assume that firms at the least *believe* AI chatbots positively impact sales.

a conversation with a salesperson. In such conversations, neither 'agent' (the customer or the salesperson) knows the exact motive of the other. These goals are likely different but may align. For instance, a customer who wants a high-quality product may be satisfied by a salesperson who sells them an expensive product, assuming a loose correlation between quality and price. Metacognition enters the fray insofar as each agent is trying to determine what the other agent's objective is, so they can adopt appropriate strategies to either persuade an agent (from the salesperson's perspective) or to defend from persuasion (from the customer's perspective).[24]

Now replace 'salesperson' with an AI chatbot. The first benefit of doing so is one may gain a new perspective on the question of trust of AI. The degree to which people trust AI is likely influenced by the context in which the AI chatbot is deployed. For instance, an (ideal) government acts in the interests of its citizens. As such, metacognitively a user might determine that regardless of the chatbot's capabilities, its objective *is to help them*, leading to higher trust. In a different context—online retail, for instance—it is reasonable to assume adversarial objectives between the chatbot and the user.[25] Indeed, even in the governmental context, this might explain why Aoki (2020) finds trust in AI chatbots to be influenced by the policy context in which the chatbot is deployed.

Where this metacognitive perspective becomes interesting is around the question of knowledge. Friestad and Wright (1994) argue in their model of metacognition that knowledge is a key driver of metacognition, and thus successful persuasion. Knowledge takes different forms. For instance, someone who is very knowledgeable about a product will likely discern the persuasion of a less-knowledgeable salesperson, as the salesperson may advocate for a product the knowledgeable consumer knows to be inferior (Moon, 2010). Another form of knowledge is knowledge of persuasion techniques themselves. Salespeople will often encourage wavering customers to think about how they would use a product once they have purchased it ('call to action'). They might recount the 'testimony' of other satisfied customers ('appeal to conformity'). A customer's knowledge or ignorance of such techniques is likely to influence they metacognition.[26] But perhaps most interestingly, agential knowledge of one another influences metacognition.

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

A model trained on a large body of text will likely be able to demonstrate superior persuasion knowledge for similar reasons to its superior product knowledge. But such a model may also be able to adapt its use of such techniques in response to ongoing dialogues with users.[27] For instance, *mirroring* is a persuasion technique which involves matching the behaviour and personality style of the opposing agent. Marketing research suggestions people respond more positively to sales

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

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

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

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

strategies which match their personality, making mirroring an effective sales technique (Moon, 2002). Murphy (2024) finds that LLMs like GPT-3.5 and GPT-4 can predict individual personalities from text data with around a 75% accuracy.[28] They note that such capabilities could be used to influence people. One immediate application might be for an AI chatbot to mirror the personality of the customer it is engaged with.

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

When communicating with an AI chatbot, users have a powerful piece of agential knowledge—they know the chatbot is, well, a chatbot. But they will likely know little else which might provide them with a strategy for metacognitive defence.[29] By contrast, AI chatbots could be designed to incorporate a huge array of agential knowledge about the user. Some knowledge might be derived—as in the case of personality—but other knowledge already be held by the retailer and put to use by the chatbot.[30] For instance, the purchase history and location of a previous customer will likely be known to a retailer. Such information might then be used to develop persuasion techniques against which a customer is more vulnerable. This is not necessarily a *negative* application—the use of superior agential knowledge of the customer might allow the customer to achieve superior outcomes—but in an adversarial, metacognitive context, such an application is likely to be deployed primarily for the maximal benefit of the retailer. For instance, if a retailer knows (approximately) when a spouse's birthday is, an AI chatbot might *naturally* incorporate into an unrelated sales chat a reminder to buy them a present. Of course, at the opportune time of the year.

Economically, it is interesting to note that AI chatbots enhance the value proposition of sales. A salesperson will likely discard much of the agential knowledge they gain about a customer once the customer leaves the store—particularly if the customer does not buy anything. An interaction with an AI chatbot, by contrast, leaves a data trail which firms can mine for greater customer insights and, ultimately, sales (Zuboff, 2015). While mere speculation, one might hypothesise that the optimal deployment of an AI chatbot (in *sales*) is not whatever allows the

[28] Murphy (2024) uses the 50 most recent tweets of people contained with the same data. Precisely how much data is needed to determine personality is an important question, given low user engagement with AI chatbots (at present), and given the likely short conversations those who use chatbots actually have with them. For instance, if very little text can accurately predict personality, than a retailer might prioritise *any* engagement over pro-longed engagement—especially if tried to a customer account, allowing a long-term profile to be assembled. If longer conversations are needed, a retailer might be incentivised to prolong conversations, leading to different prompts and designs of chatbots. If mirroring personality does not significantly increase sales, such adjustments may not be implemented. Though, this is not necessarily the only reason why a retailer might seek to retain customers in chatbot conversations.

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

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

customer to find what they are looking for promptly, nor even whatever encourages a customer to buy a more expensive product, but what allows a retailer to exact as much agential knowledge from the customer as possible while still maximising sales.[31] From an OCA perspective, which is often concerned with consumer welfare, this and above speculations raise important questions which should elevate the study of AI chatbots to the forefront of researcher's minds.

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

Chopra and Haaland (2023) champion this approach because it may allow qualitative researchers to engage in topics and fields which have typically shunned qualitative methods, like economics. Such an ambition is admirable. Nevertheless, the application of AI chatbots here is hardly perfect. For instance, one reason for small samples in qualitative research is the scarcity of suitable interviewees. For instance, there may only be a handful of experts on a topic or phenomenon in the whole world.

Digital Clones and Simulated Societies

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

At the heart of this idea is the notion of *algorithmic fidelity*. Argyle *et al.* (2023, p. 339), who propose the term, define algorithmic fidelity as, "the degree to which the complex patterns of relationships between ideas, attitudes, and sociocultural contexts within a [large language] model accurately mirror those within a range of human subpopulations." An AI model is high in fidelity if it produces outputs which correspond to those which might be produced by various smaller subgroups. Likewise, low fidelity models may arise through poor prediction of subgroups, or through an emphasis on the *average* of the whole population. Argyle (2023, p. original emphasis) contend that emerging AI models demonstrate high fidelity because, "these language models do

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

not contain just one bias, but *many*" allowing AI models to be, "biased both toward *and* against specific groups and perspectives in ways that strongly correspond with human response patterns along fine-grained demographic axes."

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

However, these studies in consumer behaviour do not reflect the full literature on silicon sampling. Studies on political decision-making in particular demonstrate various challenges related to minority representation in LLMs. Lee *et al.* (2023) find GPT-4 can generate synthetic populations which accurately simulate presidential voting behaviours and policy positions.[32] Though, such high accuracy is only demonstrated when GPT-4 is given contextual priming and psychological data about the simulated individual. Even with such additional prompting, Lee *et al.* (2023) report discrepancies, with GPT-4 under-estimating support for some policies amongst minority groups such as Black Americans. In a similar study, Hwang *et al.* (2023) report comparable results. In yet *another* comparable study, Santurkar *et al.* (2023) find poor simulation accuracy of political opinions when LLMs are asked to simulate 60 different US minority groups. Greater fine-tuning—what Santurkar *et al.* (2023, p. 29971) call "steering"—fails to improve accuracy.

For Santurkar *et al.* (2023), a major reason for poor simulation is poor training data. They suggest that popular LLMs like GPT-4 have simply not been trained on enough data originating from or representative of minority groups. A similar argument is given by Shrestha *et al.* (2024). In

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

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

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

On the other hand, under-representation in data may reflect an array of social and economic barriers to representation, which points to a more substantial methodological challenge for silicon sampling approaches (Sorensen *et al.*, 2024).[33] In such instances, under-representation reflects a *bias* against some groups, rather than merely *reflecting* their minority within society as a minority of examples within the dataset. This is a perspective which Argyle *et al.* (2023) acknowledge and believe is essential when evaluating algorithmic fidelity. Even if an LLM has "many biases," one bias by still exert an outsized influence on the final output, essentially *crowding out* other biases within the model (Sorensen *et al.*, 2024). For instance, Peterson (2024) has demonstrated that the training of LLMs typically involves the training 'long tails' *out* of the model. These long tails capture observations which deviate from the average, and which in social terms will often represent minority groups and their views.[34] Removing these outliers causes LLMs to improvement in terms of average performance, but at the expense of the average becoming *over-represented* within the model. Peterson (2024) thus argues that LLMs and similar AI systems are effective aggregators of populations but may promote an artificial consensus when deployed to *simulate* populations.[35]

Amirova *et al.* (2023) offer an additional, and very interesting, critique of silicon sampling. Focusing on the applicability of silicon sampling for *qualitative* research, Amirova *et al.* (2023) use GPT-3.5 to generate synthetic interviewees. The qualitative data which is subsequently generated was then analysed using qualitative research methods, rather than simply comparing quantitative measures (as many other silicon sampling studies do). Amirova *et al.* (2023, p. 1) find that in discussion of key themes and broad discussion topics, LLM simulations are "strikingly similar" to responses given by people whom they also interview. However, when a more detailed analysis of

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

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

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

the interviews is undertaken, synthetic interviewees differ substantially from their human counterparts, including in terms of tone, structure, and language style. Amirova *et al.* (2023) thus conclude that silicon sampling fails to demonstrate anything more than a high-level approximation of human populations.

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

Thus, with a critical eye, it is difficult to suggest that silicon sampling is at present a reliable method within behavioural science, and within the wider social sciences.[36] This is to be expected with any novel method, with important challenges around representation, measurement of accuracy, and practitioner motivation to demanding resolution. An alternative simulation method, utilising generative AI, may resolve some of these challenges associated with silicon sampling, though introduce others in turn.

Digital cloning is a broad term to describe the simulation of people using generative AI technologies. Digital clones can be created in several ways, and deployed for a multitude of ends, some more controversial than others. For instance, demographic and psychological data could be used to create a digital 'cognitive' clone of an individual. This person could then use the clone as a cognitive aid, helping them make decisions in everyday life.[37] For instance, Golovianko *et al.* (2023) argue that critical decision-makers (e.g., doctors, executives, military commanders), whose absence may have an outsized impact on the outcomes of a decision, may utilise digital clones as 'donors' when they are not available (e.g., due to sickness).[38] This builds from previous work by

[36] Though, the same critical eye may conclude that the cost savings of silicon sampling, coupled with the *appearance* of promoting inclusivity, may motivate further experimentation and deployment of the approach in the coming years (Mills and Sætra, 2024b). I will return to such speculation in Chapter 5.

[37] An appliable term here is *exogeneous cognition*. This term has been used by some in the marketing literature to describe technologies which eliminate or assist with consumer decision-making (Smith *et al.*, 2020). For example, recommendation algorithms may be considered a kind of exogeneous cognition, as might some forms of paternalist AI discussed in Chapter 2. Indeed, any technology which aids a decision-maker in navigating a decision, say through reducing information or making obscured information more accessible, can be understood as a kind of exogeneous cognition. This is to say, a valid alternative title for this chapter would be 'Exogeneous Cognition.'

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

Noting this is important because it demonstrates an important economic criticism of a simulation methods like digital cloning. Transaction cost economists argue that high human specificity is a boon for those who have it, as it allows them to demand greater compensation (as their absence is more impactful, and their skills harder to re-acquire). Digital cloning may thus be understood as a kind of *automation*, which might have a downwards pressure on

Golovianko *et al.* (2021), who suggest digital 'physical' clones could act as avatars for a person, allowing them to be in many different places at once.[39] Digital cloning has even sparked discussion around *resurrection*, in a manner of speaking. If AI technologies can accurately simulate the mannerisms and personality of individuals, then one might choose to digitally clone a deceased loved one (Iwasaki, 2023); desires for a digital 'after life' might even prompt some to digitally clone themselves *while they are alive* as a kind of autobiographical (or self-obsessed) exercise. Applications such as these remain controversial (Iwasaki, 2023).

Of course, digital clones might be used for more conventional objectives. Truby and Brown (2021) argue that digital cloning is a natural extension of digital surveillance and micro-targeting practices which characterise modern online advertising and retailing. They suggest that firms may develop digital clones of people without meaningful consent,[40] using these clones to experiment with different marketing and persuasion strategies to more effectively advertise to 'real' consumers. In many ways, such a hypothesis aligns both with discussions of personalisation in Chapter 2, and forthcoming in Chapter 4, as well as discussions of persuasive AI chatbots found in this chapter.

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

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

The sequel comes from Park *et al.* (2023). As previously, this is not a digital cloning study. Instead, this study leverages generative AI even more to simulate much more complicated agent behaviour. 25 artificial agents are created and placed in an interactive sandbox to interact with one another. These agents have the same complex set of bespoke traits which define their backgrounds. However, their interactions are simulated generatively through and LLM, with these interactions *changing* the artificial agents over time. In this sense, the agents *remember* and may simulate *planning* of their behaviours, becoming "*generative* agents" (Park *et al.*, 2023, p. 1, emphasis added). In one fascinating instance, a single agent proposed throwing a Valentine's Day party; subsequent agents then began to plan for this party, both by scheduling the party and finding fellow agents to bring to the party as dates. This leads the researchers to argue that the behaviour demonstrated in the

earnings (while having an upwards effect on productivity—this raises an interesting, but tangential, question about *who owns the digital clone*?).

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

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

simulation is meaningfully similar to human group behaviour, and much more complicated than simulation techniques lacking the generative AI component.

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

Compared to the original social attitudes survey, digital clones accurately predicted the attitudes of their social counterparts in 69% of instances. However, accounting for the rate at which participants themselves were consistent (around 81% consistency over two weeks), the normalised accuracy of the digital clones rises to around 85%. Correlations between clones and participants, in terms of personality and economic behaviour, were also relatively high at around 75%, and 60%, respectively. And, in terms of representativeness, Park *et al.* (2024) report no meaningful difference in accuracy for minority participants.[41] Though, such levels of accuracy were only found when clones were created from extensive interview data. Clones created using a simple personal description, or demographic data, did not demonstrate as high accuracy when simulating social attitudes (around 70% normalised accuracy). Interestingly, though, interviews did not outperform these 'persona' and 'demographic' clones when predicting personality or economic decision-making.[42]

The work of Park and colleagues is fascinating. Digital cloning and simulation may allow behavioural scientists to predict behaviours, both individually and as part of a group, which could offer invaluable insights. Furthermore, one can readily imagine experimenting with policy ideas in the simulation sandbox, observing how people are likely to respond to an intervention without having to *actually* experiment on people. This may reduce the risk of harm and save money (in some instances). Combining, say, Aonghusa and Michie's (2020) work on predicting policy recommendations with digital cloning may allow rapid *testing* of those predictions, and ultimately lead to better policy.[43]

---

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

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

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

Yet, one application of digital clones within behavioural science may stand out: organisational behaviour. Organisational behaviour has been little discussed in this chapter, in part because modern behavioural science tends to focus on individual behaviour (e.g., biases) and outcomes (e.g., discrete choices).[44] Thus, the recommendations which modern behavioural science prescribes (e.g., nudging) may gain less purchase within a complicated, organisational context. Though, organisations are increasingly entering into behavioural science discussions as practitioners seek both to contribute to more challenging, complex behavioural problems (Hallsworth, 2023; Sanders, Snijders and Hallsworth, 2018), and as pressure pushes the field to demonstrate relevance in increasingly challenging, complex policy areas (Chater and Loewenstein, 2022).[45]

Hallsworth (2023, p. 312) has called for behavioural science to "see the system" emphasising that behavioural science interventions do not happen in a vacuum, but instead within a context which produces ripple effects and feedback mechanisms. This call is combined with a call to bring insights from systems thinking and systems analysis into behavioural science practice.[46] One attempt to do this might be to target interventions based on their place within a hierarchy, or within a social network. For instance, influencing the behaviour of a manager might have a greater effect than a worker, as the manager's behaviour is likely to have ripple effects impacting those who work under them. Understanding such dynamics has long been of interest to behavioural scientists (Dolan and Galizzi, 2015; Sanders, Snijders and Hallsworth, 2018).[47] One may speculate about the applications of AI to such problems. For instance, a predictive AI system (or even a relatively simple algorithm) might be able to score the 'influence' of each member of an

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

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

The foundation of this idea is provided in Cyert and March's (1963) behavioural theory of the firm, which in part emphasises how different stakeholders within a firm pursue their individual goals *through* the firm, with the firm's success determined by the extent to which the means for stakeholders to achieve their goals complement, rather than contradict, one another (also see Simon's *Administrative Behaviour*). This, according to Simon (2000), is the essence of *cooperation*. Likewise, the extent to which a node 'lights up' is based, perhaps, on the extent to which nodes in the previous layer 'cooperate' and 'light up' at the same time. It is perhaps telling that Minsky (1986), in describing intelligence as a 'society' or collective of simpler information processing units, discusses democracy. Neither is it surprising that Hayek, whose *Sensory Order* (Hayek, 1999) deals with connectionist psychology, also championed the free market as an efficient, information processing unit made up of decentralised individuals each pursuing their own objectives (Hayek, 1945). This notion that smaller, simpler information processing units may, in the aggregate, achieve successful, complex, *intelligent* behaviour, is an isomorphic description which connects brains, organisations, democracies, and now, machines.

This is not to suggest this isomorphism is *all there is* to the behaviour of these information processing systems. Rather, it is to point out that organisations have a place in this conversation, both as entities which exhibit intelligent behaviour and—insofar as artificial intelligence is simply that which is not natural (human) intelligence—as a kind of *AI* (Simon, 1981). This, more than any other, is the golden thread through Herbert Simon's eclectic research career.

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

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

organisation, so that interventions may be targeted at optimal leverage points. Following Chapter 2, AI systems might be used to personalise interventions not to the individual, but to the *organisation*, insofar as such personalisation is done to manage the spread of the intervention (or the intervention's effect) throughout the organisation.

One challenge with understanding organisations, and thus successfully intervening to affect change within them, is that an organisational rarely behaves in a manner congruent with how it is, on paper, organised (Peters, 2017). One might be able to sketch out the hierarchy of an organisation and rank the 'influence' of its members by each's position within that hierarchy.[48] One might use some form of network analysis, such as looking at email correspondence, to approximate who contributes most to driving the dynamics of an organisation. But such perspectives ignore important, often undetectable interpersonal factors which might, across time and hierarchy, matter much more to the behaviour of an organisation, and its members.

Imagine a simple organisational structure consisting of person A—a senior manager—B—a junior manager—and C—an entry-level worker. On paper, A > B > C in terms of organisational authority. If one wished to affect the behaviour of B and C, targeting A would seem to make a lot of sense in this *linear* hierarchy. But organisational scholars—if not behavioural scientists—will frequently stress that members of an organisation regularly embody *several different* roles at the same time (Simon, 2000). A is not just a manager; they might be a sports fan. And C, while an entry-level worker, might support the same team as A. All the while, B supports a bitter rival. At the coffee machine or the water cooler each morning, then, while one might often observe a situation where A = C, and even where C > A. These competing maps—or *graphs*—of the organisation exist simultaneously (e.g., A > B > C *and* C >= A > B), and lead to a fundamental challenge around the question of influence—should an intervention target A, or C?

Peters (2017) resurrects a cybernetic term—*heterarchy*—to describe such problems. Heterarchy was originally proposed by Warren McCulloch (1943)—another 'father' of AI—to describe how complex behaviour arises within systems of agents. One can imagine that in an organisation described as A > B > C, complex behaviour cannot emerge. A is always the focal point; B responds to A's actions, and C to B's. Hence why, as above, this hierarchy can be described as linear. But in the second description, where C >= A > B *and* A > B > C, much more complicated behaviours can arise. For instance, B can direct C to perform a task; C can then perform the task, or attempt to influence A, who can then direct B to do something different, which then impacts C, and so on.[49] Heterarchy, then, is a view of system behaviour which incorporates both the formal

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

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

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

and informal connections between agents within the system,[50] and holds that systems "are neither ordered nor disordered but instead are ordered complexly in ways that cannot be described linearly" (Peters, 2017, p. 23). And, as Peters (2017) argues, it is often the 'informal' structure of a system which determines how it changes, and thus how the system *can be changed*.[51]

Behavioural scientists who wish to 'see the system' might be said to have embraced the idea of heterarchy insofar as there is growing discussion of 'behavioural systems mapping.' Behavioural systems mapping seeks to draw a 'map' of "the key actors in a scenario and their behaviours" and "how these behaviours relate to each other" (West *et al.*, 2020, p. 27). In the above hypothetical, the notions of A > B > C and C >= A > B could be viewed as rudimentary system maps, insofar as one can identify 'nodes' (A, B, and C) and draw arrows of influence connecting these nodes. One potential difficulty of a behavioural systems mapping approach, however, is that it is not clear how different maps can exist *simultaneously*, and thus capture a multitude of different roles which individuals of an organisation may adopt.[52] For instance, what if some tasks A is willing to overrule B's command of C, but in others, A steadfastly backs B? Which map should be used, and when? As Peters (2017, p. 23), writing on McCulloch, notes, "even the simplest systems can be subject to multiple competing regimes of evaluation."

Simulations with digital clones may resolve these problems and allow behavioural science and behavioural systems mapping to incorporate heterarchical thinking into the design and implementation of interventions. This is not to suggest that a single simulation predicts the future of a fundamentally complex (if not chaotic) system. Rather, it is to suggest that through multiple simulations of digital clones interacting with one another, one might be able to observe patterns and trends in those interactions which reflect the heterarchical structure of a real organisation; a

---

[50] One (say, an economist) might object that while such social connections are not unrealistic, A, B, and C remain ultimately bound by the authority of the organisation, and that an embrace of the informal may be mitigated through adequate organisational incentives (e.g., Jensen and Meckling, 1976). Such an objection is why it is important to incorporate bounded rationality theories of organisations (e.g., Cyert and March, 1963; Simon, 2000) into the analysis, as these perspectives emphasise—both theoretically and through observation—that sensitivity to incentives is often less than natural propensity to messy, but *humans*, interpersonal dynamics.

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

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

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

structure which may be difficult to observe through other methods, which may elude even the organisation itself; and if not, a structure which the organisation might never willingly admit is demonstrative of how it operates.

Solutions With Problems

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

The focus of this chapter has been on how AI can support behavioural science tackle the problems of too much information, and obscured or hidden information. Though, from the outset, the reader has been encouraged to understand information not in terms of *bits*, but as something with a social life. In doing so, rather than merely describing applications, this chapter has considered how similar applications could be applied to different *ends*, and how the merits of various applications become complicated when considered within different social contexts. The following chapter, Chapter 4, elaborates on many of the social implications of AI in behavioural science which this chapter has begun to unpick.