ISSN 1936-5349 (print)
ISSN 1936-5357 (online)

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS

THE OPPORTUNITIES AND COSTS OF AI IN BEHAVIOURAL SCIENCE

Stuart Mills
Samuel Costa
Cass R. Sunstein

Discussion Paper No. 1104

07/2023

Harvard Law School
Cambridge, MA 02138

The Opportunities and Costs of AI in Behavioural Science

Stuart Mills,[1] *Samuël Costa*,[2] *Cass R. Sunstein*[3]

Abstract[4]

This article discusses the opportunities and costs of AI in behavioural science. We argue that because of pattern detection capabilities, modern AI will be able to identify (1) new biases in human behaviour and (2) known biases in novel situations. AI will also allow behavioural interventions to be personalised and contextualised, and thus produce significant benefits. Finally, AI can help behavioural scientists to 'see the system,' by enabling the creation of more complex and dynamic models of human behaviour. While these opportunities will significantly advance behavioural science and offer great promise to improve the lives of citizens and consumers, we highlight several costs of using AI. We focus on some important environmental, social, and economic costs that are relevant to behavioural science and its application. Some of those costs involve privacy; others involve manipulation.

[1] Assistant Professor of Economics, University of Leeds. Available at: s.mills1@leeds.ac.uk
[2] PhD Candidate, Department of Experimental Psychology, Ghent University
[3] Robert Walmsley Univeristy Professor, Harvard University.
[4] The authors are grateful to Michael Hallsworth, Edward Bradon, Edward Flahavan, and Maximilian Kronerdale for their constructive comments and suggestions. All errors are our own.

Introduction

To say the least, artificial intelligence (AI) is developing with extraordinary speed. ChatGPT, an AI chatbot developed by OpenAI, is the fastest growing online service in history (Ahuja, 2023). The implications of AI for behavioural science may be particularly significant, extending far beyond the historic connection (Simon, 1981). Modern AI excels at pattern detection, from identifying animals within images to predicting text from an initial prompt. Modern behavioural science, particularly over the past 15 years, has focused on identifying and operationalising bias and noise in human decision-making, and to providing correctives to reduce the effects of each (Hallsworth, 2023; Hallsworth and Kirkman, 2020; Halpern, 2015; Kahneman, 2011; Kahneman, Sibony and Sunstein, 2021; Thaler and Sunstein, 2008). Bias and noise are, essentially, behavioural patterns. Thus, AI is likely to be valuable within behavioural science for modelling and examining human behaviour and perhaps for improving it, or improving on it (Mills, 2022a; Ludwig and Mullainathan, 2022). For that reason, the use of AI alongside behavioural science is likely to be widespread in many applicable domains, such as consumer research and public policy (Sunstein, 2023).

This article outlines some opportunities and costs of AI-based behavioural science, including algorithmic behavioural science, in the coming years.

We highlight important work already done to identify discriminatory biases, such as racist and sexist word associations (*d-biases*), within natural language text via AI methods (Bolukbasi *et al.*, 2016; Brunet *et al.* 2019; Caliskan *et al.*, 2017). At the same time, we note that relatively little work (Horton, 2023; Jones and Steinhardt, 2022) to date has used AI to identify cognitive biases (*c-biases*), which are the focus of modern behavioural science. This is a clear, immediate opportunity for AI in behavioural science research (Ludwig and Mullainathan, 2022, 2021; Mills, 2023; Sunstein, 2022a, 2022b, 2019).

Modern behavioural science has also received significant criticism in recent years (Chater and Loewenstein, 2022; Maier *et al.*, 2022), some of it highlighting the need for more contextualised behavioural approaches that incorporate heterogeneity (Hallsworth, 2023a, 2022; Hecht *et al.*, 2022; Mills, 2022b, 2021; Schimmelpfennig and Muthukrishna, 2023; Sunstein, 2023; Szaszi *et al.*, 2022). This 'heterogeneity revolution,' (Bryan, Tipton and Yeager, 2021) is likely to be promoted and accelerated by AI technologies (Agrawal *et al.*, 2022; Michie *et al.*, 2017; Mills, 2022a; Rauthmann, 2020), both as a new tool for behavioural science and in conjunction with existing strategies, such as mega studies (Buyalskaya *et al.*, 2023; Duckworth and Milkman, 2022; Milkman *et al.*, 2022; Milkman *et al.*, 2021).

Finally, from a complex systems perspective, AI has the potential to help behavioural scientists to 'see the system' (Hallsworth, 2023b). This may be through predicting the optimal timing and context for delivery of interventions (Mills, 2022c; Yeung, 2017). It may also take the form of probing human behaviour as a complex system to identify optimal leverage points for affecting behaviour change (Hallsworth, 2023b; Park *et al.*, 2023; Schmidt and Stenger, 2021).

AI also creates new costs for practitioners and consumers. We briefly address the environmental effects of AI in behavioural science (Crawford, 2021; Dhar, 2020; Wu *et al.*, 2022). Where behavioural science uses AI in behavioural interventions to promote pro-environmental consumer behaviours, these energy-intensive methods must factor into the final evaluation of the intervention. However, environmental costs will affect any and all disciplines that use AI. As such, we focus more on costs specific to behavioural science practitioners and consumers.

AI-behavioural models may impose substantial social costs, as by endangering privacy through data collection (Hagendorff, 2022; Sætra, 2020; Saheb, 2022), and interfering with the formation of individual preferences (Bommasani *et al.*, 2022; Russell, 2019; Sunstein, 2022a). The latter risk is particularly important when considering vulnerable individuals, such as children (Akgun and Greenhow, 2022; Smith and de Villiers-Botha, 2021). At least with regulation of various kinds, AI may be limited in its ability to accommodate important individual and societal values, and that limitation may undermine public trust and produce welfare costs from interventions otherwise forgone (Mills, 2023). Finally, AI-behavioural approaches may not be economically viable in some domains where existing behavioural science methods are appropriate (Mills, 2022b; Sunstein, 2023, 2012). Furthermore, skill premiums are likely to be high for professionals who command effective knowledge of behavioural science and AI, meaning that – at least in the near-term – established methods may prove more economically viable (Hallsworth, 2023b; Lipton and Steinhardt, 2018).

Understanding the opportunities of behavioural science and AI, as well as these costs, will be crucial for determining best-practice applications, and regulatory policy to protect consumers and citizens.

Opportunity 1: Identifying Biases

While behavioural science uses a suite of tools to affect behaviour change (Hallsworth, 2023b, 2022), and points to the need to go beyond merely identifying 'flaws' in human behaviour (Bryan, Tipton and Yeager, 2021; Gigerenzer, 2018; Nisa *et al.*, 2020; Schimmelpfennig and Muthukrishna, 2022), identifying bias and noise with AI is a clear opportunity for behavioural science. Behavioural biases can be understood as predictable patterns or error in human behaviour (Kahneman, 2011; Thaler and Sunstein, 2008, 2003; Tversky and Kahneman, 1974), and the pattern-detecting capabilities of modern AI are likely to be well-suited to the task of identifying biases from behavioural data (Kleinberg *et al.*, 2018; Kleinberg *et al.*, 2015; Ludwig and Mullainathan, 2022, 2021; Mills, 2023; Sunstein, 2022a, 2022b, 2019). In fact, AI may identify biases that have never been identified before (Ludwig and Mullainathan, 2022). Equally, noise may hide patterns in behaviour that humans may fail to spot, but that AI can identify and quantify (Aonghusa and Michie, 2020).

AI has been used to identify discriminatory biases within human behaviour. For instance, *Word2Vec* is a natural language processing AI developed by Google (Mikolov *et al.*, 2013). Like many natural language AI systems, *Word2Vec* identifies the statistical relationships between words in terms of probabilities and uses these relationships to identify word associations (Wolfram, 2023). A user can then explore these associations through posing questions to the AI. Through such questioning, *Word2Vec* has often been found to produce gender-biased word associations (Bolukbasi *et al.*, 2016; Brunet *et al.* 2019). 'Word embedding' models such as *Word2Vec* have also been used as 'Word Embedding Association Tests' (WEATs) to replicate the results of the Implicit Associations Test (IAT) using only (big) text data (Caliskan *et al.*, 2017; Evenepoel, 2022). In both instances, only natural language is used to identified various discriminatory biases, and thus it is not that the AI systems themselves are biased, but rather, that AI can be used to identify implicit biases in natural language that were previously hidden (Brunet *et al.*, 2019).

These results suggest several opportunities. Such approaches represent alternatives approaches to, say, the IAT, for investigating human behaviour. Methods such as the IAT can be challenging to implement and time-consuming (and raise questions about external validity). Furthermore, AI approaches can unlock new avenues for behavioural research. For instance, the WEAT can be applied to any corpus of natural language data and can thus be used to explore implicit biases

across different cultural groups and time periods (Evenepoel, 2022). One need not focus on language; the potential is much broader. AI pattern detection has been used to investigate the decision-making processes of judges and doctors, with practices such as 'mugshot bias' (the tendency to rely heavily on a defendant's mugshot) identified through AI analysis (Kleinberg *et al.*, 2019; Kleinberg *et al.*, 2018; Ludwig and Mullainathan, 2022, 2021; Sunstein, 2022a).

We are speaking here of discriminatory biases, or d-biases. While such biases have a long association with behavioural science, they are distinct from the cognitive biases (Wilke and Mata, 2012) – or c-biases – which generally concern modern behavioural science (Sunstein, 2022b). This is important to note to distinguish discussions of AI for detecting biases in behavioural science from the extensive literature on algorithmic bias (which generally focuses on d-biases). Relatively little work to date has explored the use of AI to identify c-biases (Horton, 2023; Jones and Steinhardt, 2022), though importantly, some AI-based analyses have shown judges (Kleinberg *et al.*, 2018; Ludwig and Mullainathan, 2022) and doctors (Mullainathan and Obermeyer, 2022) to use more prominent information in a manner which is indictive of availability bias and representativeness bias (Mills, 2023; Sunstein, 2022b). AI techniques have also been used to study habit formation behaviour within especially large datasets, identifying important factors that influence consumption habit formation, which may have been difficult to determine via traditional statistical techniques (Milkman *et al.*, 2023).

The relative paucity of such work should be seen as a compelling opportunity for research within behavioural science. Indeed, it is hardly premature to speculate about the possibilities such a research programme might hold. For instance, real-time data on the behaviour of a financial stock trader – such as the status of their portfolio, the speed of their mouse clicks, the frequency of their email communications, and so on – might be used to predict whether the broker is in a 'hot' state, and automatically trigger risk management procedures ranging from nudge-like interventions (e.g., "you should take a break from the desk") to more coercive interventions (e.g., imposition of temporary trading limits).

Opportunity 2: Integrating Heterogeneity

Beyond expanding the toolkit by which researchers investigate human behaviour, AI presents a unique opportunity for behavioural science to progress in a way that meets various concerns about the field.

Recent high-profile results have sparked considerable debate (Hallsworth, 2023a, 2022). In particular, questions have been raised about the effectiveness of some behavioural interventions (Maier *et al.*, 2022), given what are often small effect sizes (Beshears and Kosowsky, 2020; DellaVigna and Linos, 2022; van der Linden and Goldberg, 2020). Concern has also been raised about the value of behavioural interventions that are focused on individual behaviour (Chater and Loewenstein, 2022), given current policy challenges such as climate change (Bergquist *et al.*, 2023; Nisa *et al.*, 2020). These concerns supplement earlier concerns about certain uses of behavioural insights in public policy, which have been challenged for potentially undermining individual autonomy and freedom of choice (Gigerenzer, 2015; Henderson, 2014; Mitchell, 2005; Rebonato, 2014, 2012; Rizzo and Whitman, 2020, 2009; Ryan, 2018; Sugden, 2013, 2009; Veetil, 2011).

These different concerns – of being insufficiently effective and disrespectful to individuals – may or may not have force, and may be addressed by better integrating individual heterogeneity and context into behavioural science (Bryan, Tipton and Yeager, 2021; Hallsworth, 2023a, 2023b, 2022; Hecht *et al.*, 2022; Mills, 2022b, 2021; Schimmelpfennig and Muthukrishna, 2023; Sunstein, 2023; Szaszi *et al.*, 2022). The effectiveness of behavioural interventions is likely to depend on a multitude

of factors, from the precise tool chosen (a default role, a warning, a reminder, a tax, a subsidy, a mandate; Sunstein, 2023), to individual traits (Mills, 2022b; Peer *et al.*, 2020; Thunström *et al.*, 2018), to strength of preferences (de Ridder, Kroese and van Gestel, 2022) to cultural factors (Schimmelpfennig and Muthukrishna, 2023).

In recent years, behavioural studies have increasingly used moderation and mediation approaches to probe behavioural results to find and identify heterogeneous effects within a sample (Dolgopolova *et al.*, 2021; Hecht *et al.*, 2022; Jachimowicz *et al.*, 2019; Nekmat, 2020; Peer *et al.*, 2020; Thunström *et al.*, 2018) – for instance, when evaluating calorie labels (Thunström, 2019) or COVID-19 interventions (Kantorowicz-Reznichenko *et al.*, 2022; Krpan *et al.*, 2021). This can lead to a deeper understanding of the factors influencing the intervention, and thus creates opportunities for interventions to be tailored to specific environments, individuals, or policy objectives (Agrawal *et al.*, 2022; Mills, 2022b; Sunstein, 2023). More tailored interventions may also empower individuals to 'self-nudge,' reassured that such interventions are attuned to their personal preferences and objectives (Krpan and Urbaník, 2021).

While such approaches are promising, and interject much needed nuance into the evaluation of behavioural results (Bryan, Tipton and Yeager, 2021; Hallsworth, 2022; Szaszi *et al.*, 2022), approaches such as analysing the potential moderators of behavioural interventions are limited by the potentially subjective choices in how the sample is stratified to investigate the effect of, say, gender or personality. Furthermore, examining all possible combinations of heterogeneous factors on an identified effect may be too resource-intensive given current research practices, as moderators themselves may be moderated by additional factors. Indeed, for n variables being examined, an approximate estimate for the number of potential models – without prior theory – would be $n!$, or n-factorial (Hayes, 2013). The question of resource intensity is particularly pertinent as behavioural science research increasingly uses 'mega studies' to investigate interventions (Duckworth and Milkman, 2022). These studies represent a very different route to understanding heterogeneous effects by embracing the power of scale. But in doing so, they are also burdened by huge amounts of data, creating an opportunity for AI to assist in the analysis (Matz *et al.*, 2017; Milkman *et al.*¸ 2023).

AI may reduce or resolve many of the challenges brought by the added complexity of heterogeneity analysis (Lazer *et al.*, 2009). Deep learning AI systems, which dominate current AI modelling, may accommodate an essentially unlimited number of input variables in an n-length input vector. For instance, rather than examining the effect of extraversion on a consumer behaviour, and *separately* examining the effect of openness on that same behaviour, an AI approach would allow each consumer's unique personality profile to be examined holistically, leading to a predictive AI model that integrates far more heterogeneity than moderation approaches can accommodate (Kosinski *et al.*, 2013; Kosinki *et al.*, 2015; Matz *et al.*, 2017). These individual-level variables are likely to be accompanied by various other contextual variables, such as time of day or location (Benartzi, 2017; Hauser *et al.*, 2014; Hauser *et al.*, 2009; Milkman *et al.*¸ 2023), to further integrate heterogeneous factors, as many 'autonomous choice architects' already do (Hermann, 2021; Hui *et al.*, 2021; Johnson, 2021; Mills, 2022a, 2022c; Mills and Sætra, 2022; Morozovaite, 2021; Yeung, 2017).

Heterogeneity-respecting behavioural interventions, developed through AI, may lead to more effective (Mills, 2022b) and equitable (Sunstein, 2023) interventions that simultaneously address concerns about the effect size of interventions given the scale of some policy challenges (Chater and Loewenstein, 2022; Nisa *et al.*, 2020). At the same time, a new-found emphasis on context and heterogeneity may turn out to be a sufficient response to the concern that behavioural

interventions are homogeneous, one-size-fits-all strategies (Hallsworth, 2022). Interesting results are already being found. For instance, AI recommendation algorithms to personalise reading recommendations for children, accounting for their abilities and tastes, have been found to produce higher levels of reading (Agrawal *et al.*, 2022).

Opportunity 3: Handling Complexity

AI invites applied behavioural science to embrace, where relevant, the complexity inherent in real human behaviour, and points towards an understanding of behaviour as part of a complex adaptive system (Hallsworth, 2023b). In some of its forms, behavioural science has several overlaps with the fields of complexity economics (Bickley and Torgler, 2021; Foster, 2006; Rosser and Rosser, 2015; Sanbonmatsu *et al.*, 2021; Sanbonmatsu and Johnston, 2019; Simon, 1981; Spencer, 2018), which uses computational techniques to model the behaviour of many artificial agents within economic systems (Arthur, 2021), and cybernetics (DeYoung, 2015; Forrester, 1971), which examines how information and feedback drive the evolution of simple and complex systems (Beer, 2002).

Behavioural interventions do not exist outside of the environment in which behaviour occurs (Banerjee and Mitra, 2023; Sanders, Snijders and Hallsworth, 2018), and furthermore, behaviour is typically not a static exercise, but a continuous one, with behaviours occurring before and after any intervention (Dolan and Galizzi, 2015; Galizzi and Whitmarsh, 2019; Krpan, Galizzi and Dolan, 2019; Maki *et al.*, 2019; Nafziger, 2020). An opportunity for AI within behavioural science is therefore predicting the optimal environments, including time of intervention delivery and before/after spillover effects of interventions (Michie *et al.*, 2017; Mills, 2022c). For instance, generative AI may be used to model many artificial agents within an 'artificial society,' to investigate behavioural responses to an intervention within a computer 'sandbox,' prior to real-world implementation (Aher *et al.*, 2023; Argyle *et al.*, 2023; Park *et al.*, 2023). This perspective requires behaviour to be viewed not as a homogeneous, individual state, but as a dynamic, adaptive response to environmental factors (Hallsworth, 2023b; Sapolsky, 2017).

Complexity and cybernetic perspectives encourage one to understand behaviour as part of a wider system where different 'variables' within the system all represent potential opportunities to intervene and affect behaviour change (Beer, 1993, 1979, 1970; Forrester, 1971). Particularly important variables within systems have been dubbed 'leverage points,' (Abson *et al.*, 2017; Leventon, Abson and Lang, 2021; Riechers *et al.*, 2021; Schmidt and Stenger, 2021). Within a complex system, these variables have an outsized effect on the system as a whole, and from a behavioural perspective, have been offered as a valuable direction for future research to understand how behavioural interventions can be targeted to produce substantial behaviour change (Abson *et al.*, 2017; Hallsworth, 2023b; Schmidt and Stenger, 2021; West *et al.*, 2020).

Identifying such points, however, may be difficult owing to the complexity of the system. Large amounts of data are required to appropriately model a sufficiently complex system (Beer, 1993; Komaki *et al.*, 2021; Meadows, 1997; Simon, 1981). Furthermore, these systems – by their nature – tend to be difficult to reduce to effective, useable models for sustained periods of time, leading to what systems theorists have dubbed the 'dancing with systems' problem (Meadows, 2001).

AI represents a promising approach for mapping behavioural systems and identifying leverage points (Ng, 2016), which in turn may enhance the effectiveness of behavioural interventions (Hallsworth, 2023b; Sanders, Snijders and Hallsworth, 2018; Schmidt and Stenger, 2021). Again, this is due to the dual technological advantages of AI in analysing large amounts of data, and dynamically detecting patterns in data. As behavioural science develops to tackle more complex

behavioural challenges, there will be a growing need for strategies to understand complexity, and design interventions capable of responding to and leveraging such complexity effectively. AI may facilitate the interjection of more complexity into this ever more interdisciplinary field.

Costs

AI will create several costs for behavioural science practitioners, and consumers. Some costs, such as the environmental cost of building, using, and maintaining massive AI systems, are costs that all disciplines that embrace AI technologies must address (Crawford, 2021; Dhar, 2020; Wu *et al.*, 2022). For instance, the carbon cost of training an AI model for a study of publication quality has been estimated to be the equivalent of the carbon consumption of approximately two average American lifetimes, or seven average global lifetimes (Hao, 2019; Strubell *et al.*, 2018). Where, say, AI-behavioural models are used to design and implement behavioural interventions to promote pro-environmental consumption decisions, the energy cost of such models must be a factor in the overall policy assessment, changing the required effectiveness of the behavioural intervention to compensate for the deleterious effects of developing and delivering it (Mills and Whittle, 2023).

Consumers and citizens might also face costs of diverse kinds; some of them are difficult to quantify. These include costs that arise from data collection, in terms of privacy costs (Hagendorff, 2022; Sætra, 2020; Saheb, 2022), and from implementation, in terms of experiential costs (Russell, 2019; Sunstein, 2022a; Tanner, 2021) such as outcome homogenization (Bommasani *et al.*, 2022). For instance, where sensitive data are required for an AI-behavioural model to effectively function, but the rationale for using such data cannot be explained to the data subject – perhaps due to a lack of theoretical underpinning (Forde and Paganini, 2019; Gibney, 2018) – there is an ever-present risk that data is being misused and privacy unjustifiably violated. Even if justifiable, the potential benefits of AI-behavioural models, in terms of predictive capacity and welfare-enhancing behavioural interventions, should not be taken as sufficient to assume consent for data collection (Sætra, 2019). Such social costs are particularly pronounced when considering vulnerable individuals, such as children, and the potential harms that AI-behavioural models may induce through intervening to change behaviour at times of critical cognitive and personal development (Akgun and Greenhow, 2022; Russell, 2019; Smith and de Villiers-Botha, 2021).

There is also a pervasive risk of manipulation (Sunstein, 2015). AI might be used to lead people in directions that are not in their interest, perhaps by exploiting a lack of information or behavioural biases (Bar-Gill *et al.*, 2023). Indeed, pattern detection abilities could enable AI not only to personalise in a way that promotes people's welfare, but also to use their biases to their detriment. The costs along these dimensions could be high.

It is important, from a policy perspective, to retain human oversight and accountability for any costs that are incurred (Mills and Sætra, 2022). Having some 'human in the loop' is recognised in emerging AI position papers, such as in the UK (UK Centre for Data Ethics and Innovation, 2020), and is supported by research into public attitudes concerning algorithmic influence (Aoki, 2021; Ingrams *et al.*, 2021; Peppin, 2022; Kozyreva *et al.*, 2021).

While one may wish to balance social costs against the estimated welfare outcomes of more accurate or personalised interventions (Sunstein, 2012), poor theoretical underpinnings of AI-behavioural models may lead to a reliance on large datasets containing potentially sensitive behavioural details, lest the accuracy of the models be undermined. Broadly, the costs of AI-behavioural models, and the enhanced accuracy such approaches might bring (Mills, 2022b; Sunstein, 2023) should be weighed against the social and welfare costs of more generic, but less data-invasive, approaches to behaviour change.

For the foregoing reasons, AI-driven approaches may be less economical than established behavioural science approaches. While contextualising interventions and using heterogeneity analysis to respect individual autonomy are substantial opportunities, it is important to recognise that behavioural science has already contributed much to public life without using such technologies (Beshears and Kosowsky, 2020; Jachimowicz *et al.*, 2019; Sanders, Snijders and Hallsworth, 2018). Where existing behavioural science competencies can deliver adequate benefits, an AI-behavioural approach may ultimately be more costly, in both time and economic costs. The cost of skills may also be a factor. As some have argued in computer science (Lipton and Steinhardt, 2018), the lack of skilled AI researcher capacity has led to limited critical oversight in AI development, with the costs of resolving this issue tied to the economic cost of enhancing skills. While emerging fields, such as behavioural data science, appear promising, there is likely to be a persistent skill premium which keeps the costs of AI-behavioural approaches high compared to established techniques, at least in the near-term.

This highlights an important additional risk: rapid deployment of AI-behavioural models is likely to demand more in terms of skills than present capacity within behavioural science can meet (Hallsworth, 2023b), which in turn creates the possibility of mis-deployment and misuse (Mills, 2023). Patience in the development of this space, coupled with efforts to build capacity and understand the necessary safeguards for AI-behavioural models – given the potential costs involved – is likely critical to the successful implementation of AI within behavioural science, and to the development of appropriate policy guidance and consumer protections.

Conclusion

The opportunities AI presents for behavioural science are significant. AI has promise as a means of probing human behavioural data to identify new cognitive biases, or to identify known cognitive biases in novel contexts. AI may also promote the 'heterogeneity revolution' in behavioural science by allowing significantly more data to be used in the design and implementation of behavioural interventions. From a complex systems perspective, AI may be well-suited for optimising the timing and context of intervention delivery, again enhancing effectiveness, as well as probing behavioural systems as a whole to predict optimal leverage points for affecting behavioural change.

AI usage in behavioural science will also create costs. As with all disciplines, behavioural science must synthesise the environmental costs of energy-intensive AI technologies into its practice. Those behavioural interventions that seek to promote pro-environmental behaviours, such a cost is particularly pertinent. AI will also create various social costs for consumers and citizens, which behavioural science must face. These include privacy costs from collecting potentially sensitive data on individual behaviour, and the risks of AI-behavioural models interfering with vulnerable individuals. There are also several economic costs. AI-behavioural models are likely to raise the skill-requirements of behavioural science practitioners, making these approaches more expensive. Where such skills are scarce, there is also the risk that such methods are used without adequate understanding or oversight, leading to misuses and welfare costs suffered by the public. Furthermore, behavioural science can already do much without AI methods, and existing competencies should always be considered in comparison to potentially more costly alternatives.

As AI technologies develop, their potential will inevitably grow. The most productive paths forward focus on the distinctive opportunities and costs of an AI-driven behavioural science, with particular emphasis on the opportunity to learn more than ever before about both bias and noise, and to use what is learned to increase human welfare.

References

Abson, D. J., Fischer, J., Leventon, J., Newig, J., Schomerus, T., Vilsmaier, U., von Wehrden, H., Abernathy, P., Ives, C. D., Jager, N. W., Lang, D. J. (2017) 'Leverage points for sustainable transformation' *Ambio*, 46, pp. 30-39

Aher, G., Arriaga, R. I., Kalai, A. T. (2023) '*Using Large Language Models to Simulate Multiple MHumans and Replicate Human Subject Studies*' ArXiv. [Online] [Date accessed: 23/06/2023]: https://arxiv.org/pdf/2208.10264.pdf

Ahuja, A. (2023) '*Generative AI is sowing the seeds of doubt in serious science*' The Financial Times. [Online] [Date accessed: 01/03/2023]: https://www.ft.com/content/e34c24f6-1159-4b88-8d92-a4bda685a73c

Agrawal, K., Athey, S., Kanodia, A., Palikot, E. (2022) '*Personalized Recommendations in EdTech: Evidence from a Randomized Controlled Trial*' ArXiv. [Online] [Date accessed: 23/06/2023]: https://arxiv.org/pdf/2208.13940.pdf

Akgun, S., Greenhow, C. (2022) 'Artificial intelligence in education: Addressing ethical challenges in K-12 settings' *AI and Ethics*, 2, pp. 431-440

Aoki, N. (2021) 'The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment' *Computers in Human Behavior*, 114, e. 106572

Aonghusa, P. M., Michie, S. (2020) 'Artificial Intelligence and Behavioral Science Through the Looking Glass: Challenges for Real-World Application' *Annals of Behavioural Medicine*, 54, pp. 942-947

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., Wingate, D. (2023) 'Out of One, Many: Using Language Models to Simulate Human Samples' *Political Analysis*, 31(3), pp. 337-351

Arthur, W. B. (2021) 'Foundations of complexity economics' *Nature Reviews Physics*, 3, pp. 136-145

Banerjee, S., Mitra, S. (2023) 'Behavioural public policies for the social brain' *Behavioural Public Policy*, DOI: 10.1017/bpp.2023.15

Bar-Gill, O., Sunstein, C. R., Talgam-Cohen, I. (2023) 'Algorithmic Harm in Consumer Markets' *Journal of Legal Analysis* (forthcoming).

Beer, S. (2002) 'What is cybernetics?' *Kybernetes*, 31(2), pp. 209-219

Beer, S. (1993) '*Designing Freedom*' Anansi: Canada

Beer, S. (1979) '*Decision and Control*' Wiley: UK

Beer, S. (1970) 'Managing modern complexity' *Futures*, 2(3), pp. 245-257

Benartzi, S. (2017) '*The Smarter Screen: Surprising Ways to Influence and Improve Online Behavior*' Portfolio: USA

Bergquist, M., Thiel, M., Goldberg, M. H., van der Linden, S. (2023) 'Field interventions for climate change mitigation behaviors: A second-order meta-analysis' *Proceedings of the National Academy of Science*, 120(13), e. 2214851120

Beshears, J., Kosowsky, H. (2020) 'Nudging: Progress to date and future directions' *Organizational Behavior and Human Decision Processes*, 161, pp. 3-19

Bickley, S. J., Torgler, B. (2021) '*Behavioural Economics, What Have we Missed? Exploring "Classical" Behavioural Economics Roots in AI, Cognitive Psychology, and Complexity Theory'* CREMA Working Paper Series no. 2021-21. [Online] [Date accessed: 11/04/2023]: https://www.researchgate.net/profile/Benno-Torgler/publication/351736247_Behavioural_Economics_What_Have_we_Missed_Exploring_Classical_Behavioural_Economics_Roots_in_AI_Cognitive_Psychology_and_Complexity_Theory/links/60a6c68da6fdcc6d626878ee/Behavioural-Economics-What-Have-we-Missed-Exploring-Classical-Behavioural-Economics-Roots-in-AI-Cognitive-Psychology-and-Complexity-Theory.pdf

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., Kalai, A. (2016) '*Man is to Computer Programmer as Women is to Homemaker? Debiasing Word Embeddings*' [Online] [Date accessed: 25/01/2023]: https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., Liang, P. (2022) '*Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?*' ArXiv. [Online] [Date accessed: 11/04/2023]: https://arxiv.org/abs/2211.13972

Brunet, M., Alkalay-Houlihan, C., Anderson, A., Zemel, R. (2019) 'Understanding the Origins of Bias in Word Embeddings' *Proceedings of the 36th International Conference on Machine Learning*. DOI: 10.48550/arXiv.1810.03611

Bryan, C. J., Tipton, E., Yeager, D. S. (2021) 'Behavioural science is unlikely to change the world without a heterogeneity revolution' *Nature Human Behaviour*, 5, pp. 980-989

Buyalskaya, A., Ho, H., Milkman, K. L., Li, X., Duckworth, A. L., Camerer, C. (2023) 'What can machine learning teach us about habit formation? Evidence from exercise and hygiene' *Proceedings of the National Academy of Science*, 120(17), e. 2216115120

Caliskan, A., Bryson, J. J., Narayanan, A. (2017) 'Semantics derived automatically from language corpora contain human-like biases' *Science*, 356(6334), pp. 183-186

Chater, N., Loewenstein, G. (2022) 'The i-frame and the s-frame: How focusing on individual-level solutions has led behavioural public policy astray' *Behavioral and Brain Sciences*, DOI: 10.1017/s0140525X22002023

Crawford, K. (2021) 'The hidden costs of AI' *New Scientist*, 249(3327), pp. 46-49

DellaVigna, S., Linos, E. (2022) 'RCTs to Scale: Comprehensive Evidence from Two Nudge Units' *Econometrica*, 90(1), pp. 81-116

De Ridder, D., Kroese, F., van Gestel, L. (2022) 'Nudgeability: Mapping Conditions of Susceptibility to Nudge Influence' *Perspectives on Psychological Science*, 17(2), pp. 346-359

Dhar, P. (2020) 'The carbon impact of artificial intelligence' *Nature Machine Intelligence*, 2, pp. 423-425

DeYoung, C. G. (2015) 'Cybernetic Big Five Theory' *Journal of Research in Personality*' 56, pp. 33-58

Dolan, P., Galizzi, M. M. (2015) 'Like ripples on a pond: Behavioral spillovers and their implications for research and policy' *Journal of Economic Psychology*, 47, pp. 1-16

Dolgopolova, I., Toscano, A., Roosen, J. (2021) 'Different Shades of Nudges: Moderating Effects of Individual Characteristics and States on the Effectiveness of Nudges during a Fast-Food Order' *Sustainability*, 13(23), e. 13347

Duckworth, A. L., Milkman, K. L. (2022) 'A guide to megastudies' *PNAS Nexus*, 1, pp. 1-5

Evenepoel, A. (2022) '*Identification of Social Bias with the Word Embedding Association Test*' Unpublished Manuscript.

Forde, J. Z., Paganini, M. (2019) '*The Scientific Method in the Science of Machine Learning*' ArXiv. [Online] [Date accessed: 15/09/2021]: https://arxiv.org/abs/1904.10922

Forrester, J. W. (1971) 'Counterintuitive behavior of social systems' *Theory and Decision*, 2, pp. 109-140

Foster, J. (2016) 'Why is Economics Not a Complex Systems Science?' *Journal of Economic Issues*, 40(4), pp. 1069-1091

Galizzi, M. M., Whitmarsh, L. (2019) 'How to Measure Behavioural Spillovers: A Methodological Review and Checklist' *Frontiers in Psychology*, 10, DOI: 10.3389/fpsyg.2019.00342

Gibney, E. (2018) 'The scant science behind Cambridge Analytica's controversial marketing techniques' *Nature*, DOI: 10.1038/d41586-018-03880-4

Gigerenzer, G. (2018) 'The Bias Bias in Behavioral Economics' *Review of Behavioral Economics*, 5, pp. 303-336

Gigerenzer, G. (2015) 'On the Supposed Evidence for Libertarian Paternalism' *Review of Philosophy and Psychology*, 6, pp. 361-383

Hagendorff, T. (2022) 'Blind spots in AI ethics' *AI and Ethics*, 2, pp. 851-867

Hallsworth, M. (2023a) '*Misconceptions about the Practice of Behavioral Public Policy*' SSRN. [Online] [Date accessed: 10/04/2023]: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4328659

Hallsworth, M. (2023b) 'A manifesto for applying behavioural science' *Nature Human Behaviour*, 7, pp. 310-323

Hallsworth, M. (2022) '*Making Sense of the "Do Nudges Work?" Debate*' Behavioral Scientist. [Online] [Date accessed: 10/04/2023]: https://behavioralscientist.org/making-sense-of-the-do-nudges-work-debate/

Hao, K. (2019) '*Training a single AI model can emit as much carbon as five cars in their lifetimes*' MIT Technology Review. [Online] [Date accessed: 11/04/2023]: https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Hauser, J. R., Liberali, G., Braun, M. (2009) 'Website Morphing' *Marketing Science*, 28(2), pp. 201-401

Hauser, J. R., Liberali, G., Urban, G. L. (2014) 'Website Morphing 2.0: Switching Costs, Partial Exposure, Random Exit, and When to Morph' *Management Science*, 60(6), pp. 1594-1616

Hayes, A. F. (2013) '*Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression Approach*' Guilford Press: USA

Hecht, C. A., Dweck, C. S., Murphy, M. C., Yeager, D. S. (2022) 'Efficiently exploring the causal role of contextual moderators in behavioral science' *Proceedings of the National Academy of Science*, 120(1), e., 2216315120

Henderson, D. R. (2014) 'Libertarian Paternalism: Leviathan in Sheep's Clothing?' *Society*, 51, pp. 268-273

Hermann, E. (2022) 'Psychological targeting: nudge or boost to foster mindful and sustainable consumption?' *AI and Society*, DOI: 10.1007/soo146-022-01403-4

Horton, J. J. (2023) '*Large Language Models as Simulated Economics Agents: What Can We Learn from Homo Silicus*?' ArXiv. [Online]: https://arxiv.org/abs/2301.07543

Hui, B., Zhang, L., Zhou, X., Wen, X., Nian, Y. (2022) 'Personalized recommendation system based on knowledge embedding and historical behavior' *Applied Intelligence*, 52, pp. 954-966

Ingrams, A., Kaufmann, W., Jacobs, D. (2021) 'In AI we trust? Citizen perceptions of AI in government decision making' *Policy and Internet*, 14(2), pp. 390-409

Ioannidis, J. P. A. (2012) 'Why Science is Not Necessarily Self-Correcting' *Perspectives on Psychological Science*, 7(6), pp. 645-654

Ioannidis, J. P. A. (2005) 'Why Most Published Research Findings Are False' *PLOS Medicine*, 2(8), e. 124

Jachimowicz, J. M., Duncan, S., Weber, E. U., Johnson, E. J. (2019) 'When and why defaults influence decisions: a meta-analysis of default effects' *Behavioural Public Policy*, 3(2), pp. 159-186

Johnson, E. (2021) '*How Netflix's Choice Engine Drives Its Business'* Behavioral Scientist. [Online] [Date accessed: 10/04/2023]: https://behavioralscientist.org/how-the-netflix-choice-engine-tries-to-maximize-happiness-per-dollar-spent_ux_ui/

Jones, E., Steinhardt, J. (2022) '*Capturing Failures of Large Language Models via Human Cognitive Biases'* ArXiv. [Online] [Date accessed: 23/06/2023]: https://arxiv.org/abs/2202.12299

Kahneman, D. (2011) '*Thinking, Fast and Slow*' Penguin Books: UK

Kahneman, D., Sibony, O., Sunstein, C. R. (2021) '*Noise: A Flaw in Human Judgement*' Little & Brown: USA

Kantorowicz-Reznichenko, E., Kantorowicz, J., Wells, L. (2022) 'Can vaccination intentions against COVID-19 be nudged?' *Behavioural Public Policy*, DOI: 10.1017/bpp.2022.20

Kim, D. A., Hwong, A. R., Stafford, D., Hughes, A. D., O'Malley, J. A. Fowler, J. H., Christakis, N. A. (2015) 'Social network targeting to maximise population behaviour change: a cluster randomised controlled trial' *The Lancet*, 386(9989), pp. 145-153

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S., (2018) 'Human Decisions and Machine Predictions' *The Quarterly Journal of Economics*, 133(1), pp. 237-293

Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z. (2015) 'Prediction Policy Problems' *American Economic Review*, 105(5), pp. 491-495

Kleinberg, J., Ludwig, J., Mullainathan, S., Sunstein, C. R. (2019) 'Discrimination in the Age of Algorithms' *Journal of Legal Studies*, 10, pp. 113-174

Komaki, A., Kodaka, A., Nakamura, E., Ohno, Y., Kohtake, N. (2021) 'System Design Canvas for Identifying Leverage Points in Complex Systems: A Case Study of the Agricultural System Models, Cambodia' *Proceedings of the Design Society*, 1, pp. 2901-2910

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., Stillwell, D. (2015) 'Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines' *American Psychologist,* 70(6), pp. 543-556

Kosinski, M., Stillwell, D., Graepel, T. (2013) 'Private traits and attributes are predictable from digital records of human behavior' *Proceedings of the National Academy of Science*, 110(15), pp. 5802-5805

Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., Herzog, S. M. (2021) 'Public attitudes towards algorithmic personalization and use of personal data online: evidence from Germany, Great Britain, and the United States' *Humanities and Social Sciences Communications*, 8(117), DOI: 10.1057/s41599-021-00787-w

Krpan, D., Galizzi, M. M., Dolan, P. (20190 'Looking at Spillovers in the Mirror: Making a Case for 'Behavioural Spillunders' *Frontiers in Psychology*, 10, DOI: 10.3389/fpsyg.2019.01142

Krpan, D., Makki, F., Saleh, N., Brink, S. I., Klauznicer, H. V. (2020) 'When behavioural science can make a difference in times of COVID-19' *Behavioural Public Policy*, 5(2), pp. 153-179

Krpan, D., Urbaník, M. (2021) 'From libertarian paternalism to liberalism: behavioural science and policy in an age of new technology' *Behavioural Public Policy*, DOI: 10.1017/bpp.2021.40

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., van Alstyne, M. (2009) 'Computational Social Science' *Science*, 323(5915), pp. 721-723

Leventon, J., Abson, D. J., Lang, D. J. (2021) 'Leverage points for sustainability transformations: nine guiding questions for sustainability science and practice' *Sustainability Science*, 16, pp. 721-726

Lipton, Z. C., Steinhardt, J. (2018) '*Troubling Trends in Machine Learning Scholarship*' ArXiv. [Online] [Date accessed: 16/09/2021]: https://arxiv.org/abs/1807.03341

Ludwig, J. Mullainathan, S., (2022) '*Algorithmic Behavioral Science: Machine Learning as a Tool for Scientific Discovery*' Chicago Booth Working Paper no. 22-15. [Online] [Date accessed: 01/03/2023]: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4164272

Ludwig, J., Mullainathan, S. (2021) 'Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System' *Journal of Economic Perspectives*, 35(4), pp. 71-96

Maier, M., Bartoš, F., Stanley, T. D., Wagenmakers, E. (2022) 'No evidence for nudging after adjusting for publication bias' *Proceedings of the National Academy of Science*, 119(31), e. 2200300119

Maki, A., Carrico, A. R., Raimi, K. T., Truelove, H. B., Araujo, B., Yeung, K. L. (2019) 'Meta-analysis of pro-environmental behaviour spillover' *Nature Sustainability*, 2, pp. 307-315

Matz, S. C., Kosinski, M., Nave, G., Stillwell, D. J. (2017) 'Psychological targeting as an effective approach to digital mass persuasion' *Proceedings of the National Academy of Science*, 114(28), pp. 12714-12719

Meadows, D. (2001) 'Dancing with Systems' *Whole Earth*, 106, pp. 58-63

Meadows, D. (1997) 'Leverage points: Places to intervene in a system' *Whole Earth*, 91(1), pp. 78-84

Michie, S., Thomas, J., Johnston, M., Aonghusa, P. M., Shawe-Taylor, J., Kelly, M. P., Deleris, L. A., Finnerty, A. N., Marques, M. M., Norris, E., O'Mara-Eves, A., West, R. (2017) 'The Human Behaviour-Change Project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation' *Implementation Science*, 12(121), DOI: 10.1186/s13012-017-0641-5

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013) '*Efficient Estimation of Word Representations in Vector Space*' ArXiv. [Online] [Date accessed: 04/07/2021]: https://arxiv.org/pdf/1301.3781.pdf

Milkman, K. L., Gandhi, L., Patel, M. S., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Rothschild, J., Bogard, J. E., Brody, I., Chabris, C. F., Chang, E. Chapman, G. B., Dannals, J. E., Goldstein, N. J., Goren, A., Hershfield, H., Hirsch, A., Hmurovic, J., Horn, S., Karlan, D. S., Kristal, A. S., Lamberton, C., Meyer, M. N., Oakes, A. H., Schweitzer, M. E., Shermohammed, M., Talloen, J., Warren, C., Whillans, A., Yadav, K. N., Zlatev, J. J., Berman, R., Evans, C. N., Ladhania, R., Ludwig, J., Mazar, N., Mullainathan, S., Snider, C. K., Spiess, J., Tsukayama, E., Ungar, L., van den Bulte, C., Volpp, K. G., Duckworth, A. L. (2022) 'A 680.000-person megastudy of nudges to encourage vaccination in pharmacies' *Proceedings of the National Academy of Science*, 119(6), e. 2115126119

Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., Park, Y., Rai, A., Bazerman, M., Beshears, J., Bonacorsi, L., Camerer, C., Chang, E., Chapman, G., Cialdini, R., Dai, H., Eskreis-Winkler, L., Fishbach, A., Gross, J. J., Horn, S., Hubbard, A., Jones, S. J., Karlan, D., Kautz, T., Kirgios, E., Klusowski, J., Kristal, A., Ladhania, R., Loewenstein, G., Ludwig, J., Mellers, B., Mullainathan, S., Saccardo, S., Spiess, J., Suri, G., Talloen, J. H., Taxer, J., Trope, Y., Ungar, L., Volpp, K. G., Whillans, A., Zinman, J., Duckworth, A. L. (2021) 'Megastudies improve the impact of applied behavioural science' *Nature*, 600, pp. 478-483

Mills, S. (2023) '*The Misuse of Algorithms in Society*' SSRN. [Online] [Date accessed: 21/04/2023]: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4400026

Mills, S. (2022a) '*AI for Behavioural Science*' CRC Press: UK

Mills, S. (2022b) 'Personalized Nudging' *Behavioural Public Policy*, 6(1), pp. 150-159

Mills, S. (2022c) 'Finding the 'nudge' in hypernudge' *Technology in Society*, 71, e. 102117

Mills, S. (2021) '*The Future of Nudging Will Be Personal*' Behavioral Scientist. [Online] [Date accessed: 10/04/2023]: https://behavioralscientist.org/the-future-of-nudging-will-be-personal/

Mills, S., Sætra, H. S. (2022) 'The autonomous choice architect' *AI and Society*, DOI: 10.1007/s00146-022-01486-z

Mills, S., Whittle, R. (2023) 'Seeing the nudge from the trees: The 4S framework for evaluating nudges' *Public Administration*, DOI: 10.1111/padm.12941

Mitchell, G. (2005) 'Libertarian Paternalism is an Oxymoron' *Northwestern University Law Review*, 99(3), pp. 1245-1277

Morozovaite, V. (2021) 'Two sides of the digital advertising coin: putting hypernudging into perspective' *Market and Competition Law Review*, 5(2), pp. 105-145

Mullainathan, S., Obermeyer, Z. (2022) 'Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care' *The Quarterly Journal of Economics*, 137(2), pp. 679-727

Nafziger, J. (2020) 'Spillover effects of nudges' *Economics Letters*, 190, e. 109086

Ng, C. F. (2016) 'Behavioral Mapping and Tracking' in Gifford, R. (eds.) '*Research Methods for Environmental Psychology*' (2016). DOI: 10.1002/9781119162124.ch3

Nekmat, E. (2020) 'Nudge Effect of Fact-Check Alerts: Source Influence and Media Skepticism on Sharing of News Misinformation in Social Media' *Social Media and Society*, 6(1), DOI: 10.1177.2056305119897322

Nisa, C. F., Sasin, E. M., Faller, D. G., Schumpe, B. M., Belanger, J. J. (2020) 'Reply to: Alternative meta-analysis of behavioural interventions to promote action on climate change yields different conclusions' *Nature Communications*, 11, pp. 3901

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., Bernstein, M. S. (2023) '*Generative Agents: Interactive Simulacra of Human Behavior*' ArXiv. [Online] [Date accessed: 20/04/2023]: https://arxiv.org/pdf/2304.03442.pdf

Pedersen, T., Johansen, C. (2020) 'Behavioural artificial intelligence: an agenda for systematic empirical studies of artificial inference' *AI and Society*, 35(3), pp. 519-532

Peer, E., Egelman, S., Harbach, M., Malkin, N., Mathur, A., Frik, A. (2020) 'Nudge me right: Personalizing online security nudges to people's decision-making styles' *Computers in Human Behavior*, 109, e. 106347

Peppin, A. (2022) '*Who cares what the public think?*' The Ada Lovelace Institute. [Online] [Date accessed: 21/04/2023]: https://www.adalovelaceinstitute.org/evidence-review/public-attitudes-data-regulation/

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., Roberts, M. E., Shariff, A., Tenenbaum, J. B., Wellman, M. (2019) 'Machine Behaviour' *Nature*, 568, pp. 477–486

Rauthmann, J. F. (2020) 'A (More) Behavioural Science of Personality in the Age of Multi-Modal Snesing, Big Data, Machine Learning, and Artificial Intelligence' *European Journal of Personality*, 34(5), pp. 593-598

Rebonato, R. (2014) 'A Critical Assessment of Libertarian Paternalsim' *Journal of Consumer Policy*, 37, pp. 357-396

Rebonato, R. (2012) '*Taking Liberties: A Critical Examination of Libertarian Paternalism*' Palgrave Macmillan: UK

Riechers, M., Loos, J., Balázsi, A., García-Llorente, M., Bieling, C., Burgos-Ayala, A., Chakroun, L., Mattijssen, T. J. M., Muhr, M. M., Pérez-Ramírez, I., Raatikainen, K. J., Rana, S., Richardson, M., Rosengren, L., West, S. (2021) 'Key advantages of the leverage points perspective to shape human-nature relations' *Ecosystems and People*, 17(1), pp. 205-214

Rizzo, M. J., Whitman, G. (2020) '*Escaping Paternalism: Rationality, Behavioral Economics, and Public Policy*' Cambridge University Press: UK

Rizzo, M. J., Whitman, G. (2009) 'The knowledge problem of new paternalism' *BYU Law Review*, 4, pp. 905-967

Rosser, J. B., Rosser, M. V. (2015) 'Complexity and behavioral economics' *Nonlinear Dynamics, Psychology, and Life Sciences*, 19(20, pp. 201-226

Russell, S. J. (2019) '*Human Compatible: AI and the Problem of Control*' Penguin Books: UK

Ryan, S. (2018) 'Libertarian paternalism is hard paternalism' *Analysis*, 78(1), pp. 65-73

Sætra, H. S. (2020) 'Privacy as an aggregate public good' *Technology in Society*, 63, e. 101422

Saheb, T. (2022) 'Ethically contentious aspects of artificial intelligence surveillance: a social science perspective' *AI and Ethics*, DOI: 10.1007/s43681-022-00196-y

Sanbonmatsu, D. M., Cooley, E. H., Butner, J. E. (2021) 'The Impact of Complexity on Methods and Findings in Psychological Science' *Frontiers in Psychology*, 11, DOI: 10.3389/fpsyg.2020.580111

Sanbonmatsu, D. M., Johnston, W. A. (2019) 'Redefining Science: The Impact of Complexity on Theory Development in Social and Behavioral Research' *Perspectives on Psychological Science*, 14(4), pp. 672-690

Sanders, M., Snijders, V., Hallsworth, M. (2018) 'Behavioural science and policy: where are we now and where are we going?' *Behavioural Public Policy*, 2(2), pp. 144-167

Sapolsky, R. (2017) '*Behave: The Biology of Humans at our Best and Worst*' Penguin Books: UK

Schimmelpfennig, R., Muthukrishna, M. (2023) 'Cultural evolutionary behavioural science in public policy' *Behavioural Public Policy*, DOI: 10.1017/bpp.2022.40

Schmidt, R., Stenger, K. (2021) 'Behavioral brittleness: the case for strategic behavioral public policy' *Behavioural Public Policy*, DOI: 10.1017/bpp.2021.16

Simon, H. A. (1981) '*The Sciences of the Artificial*' 2nd edition. MIT Press: USA

Smith, J., de Villiers-Botha, T. (2021) 'Hey, Google, leave those kids alone: Against hypernudging children in the age of big data' *AI and Society*, DOI: 10.1007/s00146-021-01314-w

Spencer, N. (2018) 'Complexity as an opportunity and challenge for behavioural public policy' *Behavioural Public Policy*, 2(2), pp. 227-234

Strubell, E., Verga, P., Andor, D., Weiss, D., McCallum, A. (2018) '*Linguistically-Informed Self-Attention for Semantic Role Labelling*' ArXiv. [Online] [Date accessed: 11/04/2023]: https://arxiv.org/abs/1804.08199

Sugden, R. (2013) 'The Behavioural Economist and the Social Planner: To Whom Should Behavioural Welfare Economics be Addressed?' *Inquiry*, 56(5), pp. 519-538

Sugden, R. (2009) 'On Nudging: A Review of *Nudge: Improving Decisions About Health, Wealth and Happiness* by Richard H. Thaler and Cass R. Sunstein' *International Journal of Economics and Business*, 16(3), pp. 365-373

Sunstein, C. R. (2023) 'The use of algorithms in society' *The Review of Austrian Economics*, DOI: 10/1007/s11138-023-00625-z

Sunstein, C. R. (2022a) 'The distributional effects of nudges' *Nature Human Behaviour*, 6, pp. 9-10

Sunstein, C. R. (2022b) 'Governing by Algorithm? No Noise and (Potentially) Less Bias' *Duke Law Journal*, 71(6), pp. 1175-1205

Sunstein, C. R. (2019) 'Algorithms, Correcting Biases' *Social Research: An International Quarterly*, 86(2), pp. 499-511

Sunstein, C. R. (2015) '*The Ethics of Influence*' Cambridge University Press: Cambridge

Sunstein, C. R. (2012) '*Impersonal Default Rules vs. Active Choices vs. Personalized Default Rules: A Triptych*' [Online] [Date accessed: 11/04/2023]: https://dash.harvard.edu/bitstream/handle/1/9876090/decidingbydefault11_5.pdf?sequence=1

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczél, B., Goldstein, D. G., Yeager, D. S., Tipton, E. (2022) 'No reason to expect large and consistent effects of nudge interventions' *Proceedings of the National Academy of Science*, 119(31), e. 2200732119

Tanner, G. (2021) '*The Hours Have Lost Their Clock: The Politics of Nostalgia*' Repeater Books: UK

Thaler, R. H., Sunstein, C. R. (2003) 'Libertarian Paternalism' *American Economic Review*, 93, pp. 175-179

Thunström, L. (2019) 'Welfare effects of nudges: The emotional tax of calorie menu labeling' *Judgment and Decision Making*, 14(1), pp. 11-25

Thunström, L., Gilbert, B., Jones-Ritten, C. (2018) 'Nudges that hurt those already hurting – distributional and unintended effects of salience nudges' *Journal of Economic Behavior and Organization*, 153, pp. 267-282

Tierney, W., Hardy, J. H., Ebersole, C. R., Leavitt, K., Viganola, D., Clemente, E. G., Gordon, M., Dreber, A., Johannesson, M., Pfeiffer, T., Hiring Decisions Forecasting Collaboration, Uhlmann, E. L. (2020) 'Creative destruction in science' *Organizational Behavior and Human Decision Processes*, 161, pp. 291-309

Tversky, A., Kahneman, D. (1974) 'Judgment under Uncertainty: Heuristics and Biases' *Science*, 185(4157), pp. 1124-1131

UK Centre for Data Ethics and Innovation (2020) '*Review into bias in algorithmic decision-making*' UK Government. [Online] [Date accessed: 21/04/2023]: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957259/Review_into_bias_in_algorithmic_decision-making.pdf

Van der Linden, S., Goldberg, M. H. (2020) 'Alternative meta-analysis of behavioral interventions to promote action on climate change yields different conclusions' *Nature Communications*, 11, pp. 3915

Veetil, V. P. (2011) 'Libertarian paternalism is an oxymoron: an essay in defence of liberty' *European Journal of Law and Economics*, 31, pp. 321-334

West, R., Michie, S., Chadwick, P., Atkins, L., Lorencatto, F. (2020) '*Achieving behaviour change: A guide for national government'* Public Health England. [Online] [Date accessed: 24/04/2023]: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/933328/UFG_National_Guide_v04.00__1___1_.pdf

Wilke, A., Mata, R. (2012) '*Cognitive Bias*' in Wilke, A., Mata, R. (eds.) '*Encyclopaedia of Human Behaviour*' 2nd ed. (2012).

Wolfram, S. (2023) '*What is ChatGPT Doing… and Why Does It Work?'* [Online] [Date accessed: 17/02/2023]: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

Wu. C., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, A., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee., H. S., Akyildiz, B., Balandat, M., Spisak, J., Jain, R., Rabbat, M., Hazelwood, K. (2022) '*Sustainable AI: Environmental Implications, Challenges, and Opportunities*' ArXiv. [Online] [Date accessed: 30/04/2023]: https://arxiv.org/pdf/2111.00364.pdf

Yeung, K. (2017) 'Hypernudge: Big Data as a mode of regulation by design' *Information, Communication and Society*, 1, pp. 118-136