Journal of Consumer Policy (2023) 46:387–400
https://doi.org/10.1007/s10603-023-09547-6

SHORT COMMUNICATION

AI, Behavioural Science, and Consumer Welfare

S. Mills[1] · S. Costa[2] · C. R. Sunstein[3]

Received: 27 June 2023 / Accepted: 10 July 2023 / Published online: 24 July 2023

Abstract
This article discusses the opportunities and costs of AI in behavioural science, with particular reference to consumer welfare. We argue that because of pattern detection capabilities, modern AI will be able to identify (1) new biases in consumer behaviour and (2) known biases in novel situations in which consumers find themselves. AI will also allow behavioural interventions to be personalised and contextualised and thus produce significant benefits for consumers. Finally, AI can help behavioural scientists to "see the system," by enabling the creation of more complex and dynamic models of consumer behaviour. While these opportunities will significantly advance behavioural science and offer great promise to improve consumer outcomes, 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. For consumers, some of those costs involve privacy; others involve manipulation of choices.

Keywords Artificial intelligence · Behavioural science · Consumer welfare · Personalisation · Algorithmic harm

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). For consumers, the most important point is that 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 consumer and investor decision-making and on providing correctives to reduce the effects of each (Halpern 2015; Kahneman et al., 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 consumer behaviour and perhaps for improving it or improving on it (Ludwig & Mullainathan, 2022). For that reason,

✉ S. Mills
s.mills1@leeds.ac.uk

[1] University of Leeds, Leeds, UK

[2] Department of Experimental Psychology, Ghent University, Ghent, Belgium

[3] Harvard University, Cambridge, USA

the use of AI alongside behavioural science is likely to be widespread in many applicable domains, such as consumer research and consumer policy (Sunstein, 2023).

This article outlines some opportunities and costs of AI-based behavioural science, including algorithmic behavioural science, in the coming years. We emphasize the benefits and costs for consumers, though occasionally we venture more broadly, and some implications for consumer policy.

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 & Steinhardt, 2022) to date has used AI to identify cognitive biases (c-biases), which are the focus of modern behavioural science. Both d-biases and c-biases matter to consumers and in many domains. This is a clear, immediate opportunity for AI in behavioural science research (Ludwig & Mullainathan, 2021, 2022; Sunstein, 2023).

Modern behavioural science has also received significant criticism in recent years (Chater & Loewenstein, 2022; Maier et al., 2022), some of it highlighting the need for more contextualised behavioural approaches that incorporate heterogeneity (Mills, 2022a; Szaszi et al., 2022). For consumers, this "heterogeneity revolution," (Bryan et al., 2021) is likely to be promoted and accelerated by AI technologies (Michie et al., 2017; 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 & Milkman, 2022).

Finally, from a complex systems perspective, AI has the potential to help behavioural scientists to "see the system" (Hallsworth, 2023). This may be through predicting the optimal timing and context for delivery of interventions designed to improve consumer welfare (Mills, 2022b; Yeung, 2017). It may also take the form of probing consumer behaviour as a complex system to identify optimal leverage points for affecting behaviour change (Park et al., 2023; Schmidt & 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). 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 consumer privacy through data collection (Hagendorff, 2022; Sætra, 2020; Saheb, 2022) and interfering with the formation of consumer preferences (Bommasani et al., 2022; Russell, 2019). The latter risk is particularly important when considering vulnerable individuals, such as children and teenagers (Akgun & Greenhow, 2022; Smith & 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. This is assuming that AI and behavioural science are used to promote consumer welfare. AI-behavioural models may be manipulative (Hacker, 2021; Sunstein, 2015) and induce harms through exploiting consumer biases (Bar-Gill et al., 2023; de Marcellis-Warin et al., 2022), contributing to the ongoing challenge of dark patterns in online consumer spaces (Helberger et al., 2022; Mathur et al., 2019). Finally, AI-behavioural approaches may not be economically viable in some domains where existing behavioural science methods are appropriate (Sunstein, 2012, 2023). 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, 2023; Lipton & Steinhardt, 2018).

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

Opportunity 1: Identifying Biases

Identifying bias and noise with AI is a clear opportunity for behavioural science. Behavioural biases can be understood as predictable patterns or errors in human behaviour, including consumption choices (Kahneman, 2011; Thaler and Sunstein, 2003, 2008), and the pattern-detecting capabilities of modern AI are likely to be well-suited to the task of identifying consumer biases from behavioural data (Kleinberg et al., 2015, 2018; Ludwig & Mullainathan, 2021, 2022). In fact, AI may identify biases that have never been identified before (Ludwig & Mullainathan, 2022). Equally, noise may hide patterns in behaviour that humans may fail to spot, but that AI can identify and quantify (Aonghusa & Michie, 2020). There are profit-making opportunities here for companies seeking to increase business; there are also opportunities, profit-making or not, to improve consumer welfare. One question is whether biases might be identified that are currently unknown.

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 identify 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 are evidently relevant to consumer behaviour, and they suggest several opportunities. Such approaches represent alternative 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 that bear directly or indirectly on consumption. 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., 2018, 2019; Ludwig

& Mullainathan, 2022). Similar biases might well be observed in consumers, though research is now at a very preliminary stage.

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 & Mata, 2012)—or c-biases—which generally concern modern behavioural science, especially in the domain of consumer and investor behaviour (Sunstein, 2022b). This is important to note so as 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 & Steinhardt, 2022), though importantly, some AI-based analyses have shown judges (Kleinberg et al., 2018; Ludwig & Mullainathan, 2022) and doctors (Mullainathan & Obermeyer, 2022) to use more prominent information in a manner which is indictive of availability bias and representativeness bias (Tversky & Kahneman, 1974). 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 (Buyalskaya et al., 2023). The potential of AI to identify the conditions under which individuals form consumption habits has immediate and obvious implications for consumers and scholars of consumption behaviour.

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, and the frequency of their email communications—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). The behaviour of consumers might similarly be tracked at relevant times and over short or long periods.

Opportunity 2: Integrating Heterogeneity

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

Recent high-profile results have sparked considerable debate (Hallsworth, 2023). 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 & Kosowsky, 2020; DellaVigna & Linos, 2022). Concern has also been raised about the value of behavioural interventions that are focused on individual behaviour (Chater & Loewenstein, 2022), given current policy challenges that involve consumers, in domains such as health, safety, and the environment (Nisa et al., 2020). These concerns supplement earlier concerns about certain uses of behavioural insights in consumer policy, which have been challenged for potentially undermining individual autonomy and freedom of choice (e.g., Gigerenzer, 2015; Rebonato, 2014).

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 theory and practice (Bryan et al., 2021; Hecht et al., 2022;

Szaszi et al., 2022). For consumers, the effectiveness of behavioural interventions is likely to depend on a multitude of factors, from the precise tool chosen (a default rule, a warning, a reminder, a tax, a subsidy, and a mandate) to individual traits (Peer et al., 2020; Thunström et al., 2018), to strength of preferences (de Ridder et al., 2022) and cultural factors (Schimmelpfennig & 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)—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, 2022a; Sunstein, 2022a). More tailored interventions may also empower consumers to "self-nudge," reassured that such interventions are attuned to their preferences and objectives (Krpan & Urbaník, 2021).

While such approaches are promising and interject much needed nuance into the evaluation of behavioural results (Bryan et al., 2021; 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 (Mills & Whittle, 2023). 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, some of it involving consumer behaviour, increasingly uses "mega studies" to investigate interventions (Duckworth & 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 (Buyalskaya et al., 2023; Matz et al., 2017).

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; 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 (Buyalskaya et al., 2023; Hauser et al., 2009, 2014), to further integrate heterogeneous factors, as many "autonomous choice architects" already do (Hermann, 2023; Mills & Sætra, 2022; Morozovaite, 2021; Yeung, 2017).

Heterogeneity-respecting behavioural interventions, developed through AI, may lead to more effective (Agrawal et al., 2022) and equitable (Sunstein, 2022a) interventions that simultaneously address concerns about the effect size of interventions given the scale of some consumer policy challenges (Chater & 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 for consumers and others; behavioural interventions are homogeneous, one-size-fits-all strategies. 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, 2023). At least in some of its forms, behavioural science has several overlaps with the fields of complexity economics (Sanbonmatsu & Johnston, 2019; Sanbonmatsu et al., 2021; 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, 1970).

Behavioural interventions do not exist outside of the environment in which consumer behaviour occurs (Banerjee & Mitra, 2023), and furthermore, such behaviour is typically not a static exercise, but a continuous one, with behaviours occurring before and after any intervention (Dolan & Galizzi, 2015; Krpan et al., 2019). An opportunity for AI within behavioural science is therefore predicting the optimal environments for consumers, including time of intervention delivery and before/after spillover effects of interventions (Michie et al., 2017). 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, 2023; Sapolsky, 2017). Recent studies have begun to investigate the suitability of these methods in consumer and marketing research (Brand et al., 2023).

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; Forrester, 1971). Particularly important variables within systems have been dubbed "leverage points" (Abson et al., 2017; Leventon et al., 2021; Schmidt & Stenger, 2021). Within a complex system model of consumers, 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, 2023; West et al., 2020) and influence consumption behaviours.

Identifying such points for consumers may be difficult owing to the complexity of the system. Large amounts of data are required to appropriately model a sufficiently complex system (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, 2023; Schmidt & 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. The result may be far more clarity, in the domain of consumer behaviour, about what works and what does not, and about what lasts and what does not.

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). 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 intervention assessment, changing the required effectiveness of the behavioural intervention to compensate for the deleterious effects of developing and delivering it (Mills & Whittle, 2023).

Consumers 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, 2023) such as outcome homogenisation (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 & Paganini, 2019; Gibney, 2018)—there is an ever-present risk that data are 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, 2020). Such social costs are particularly pronounced when considering vulnerable consumers, such as children and teenagers, and the potential harms that AI-behavioural models may induce through intervening to change behaviour at times of critical cognitive and personal development (Akgun & Greenhow, 2022; Russell, 2019; Smith & de Villiers-Botha, 2021).

For consumers, there is also a pervasive risk of manipulation (Hacker, 2021; Sunstein, 2015). AI might be used to lead consumers in directions that are not in their interest, perhaps by exploiting a lack of information or behavioural biases (Bar-Gill et al., 2023; de Marcellis-Warin et al., 2022). Indeed, pattern detection abilities could enable AI not only to personalise in a way that promotes consumers' welfare but also to use their biases to their detriment. This could exacerbate ongoing challenges surrounding dark patterns—online, deceptive behavioural practices designed to exploit consumers (Helberger et al., 2022; Mathur et al., 2019). The costs along these dimensions could be high. At the same time, AI technologies coupled with pro-consumer behavioural science could, in fact, emerge as a substantial bulwark against such abuses, providing consumers with fast, personalised information and advice (Micklitz & Pałka, 2017; Thorun & Diels, 2020).

It is important, from a consumer policy perspective, to retain human oversight and accountability for any costs that are incurred (Mills & 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 & Innovation, 2020), and is supported by research into public attitudes concerning algorithmic influence (Aoki, 2021; Ingrams et al., 2021; 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, 2022a; Sunstein, 2022a) 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 and consumer protection without using such technologies (Beshears & Kosowsky, 2020; Jachimowicz et al., 2019; Sanders et al., 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 & 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, 2023), which in turn creates the possibility of mis-deployment and misuse. Mistakes and harms to consumers are a possible consequence. 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 consumer policy guidance and consumer protections.

Conclusion

The opportunities AI presents for behavioural science are significant. For consumers, 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 designed to improve consumer welfare. 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 promoting relevant goals for consumers.

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 of diverse kinds, 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 will be used without adequate understanding or oversight, leading to misuses and welfare costs suffered by consumers. 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 consumer welfare.

Acknowledgements 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.

Declarations

Conflict of Interest The authors declare no competing interests.

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*, 30–39. https://doi.org/10.1007/s13280-016-0800-y

Agrawal, K., Athey, S., Kanodia, A., & Palikot, E. (2022). *Personalized recommendations in EdTech: evidence from a randomized controlled trial.* ArXiv https://arxiv.org/pdf/2208.13940.pdf. Accessed 23 June 2023.

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). *Using large language models to simulate multiple humans and replicate human subject studies.* ArXiv at https://arxiv.org/pdf/2208.10264.pdf. Accessed 23 June 2023.

Ahuja, A. (2023). Generative AI is sowing the seeds of doubt in serious science. *The Financial Times*. https://www.ft.com/content/e34c24f6-1159-4b88-8d92-a4bda685a73c. Accessed 1 Mar 2023.

Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics, 2*, 431–440. https://doi.org/10.1007/s43681-021-00096-7

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. https://doi.org/10.1016/j.chb.2020.106572

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*, 942–947. https://doi.org/10.1093/abm/kaaa095

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), 337–351. https://doi.org/10.1017/pan.2023.2

Arthur, W. B. (2021). Foundations of complexity economics. *Nature Reviews Physics, 3*, 136–145. https://doi.org/10.1038/s42254-020-00273-3

Banerjee, S., & Mitra, S. (2023). *Behavioural public policies for the social brain*. Advance online publication. https://doi.org/10.1017/bpp.2023.15

Bar-Gill, O., Sunstein, C. R., & Talgam-Cohen, I. (2023). *Algorithmic harm in consumer markets*. SSRN at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4321763. Accessed 23 June 2023.

Beer, S. (1970). Managing modern complexity. *Futures, 2*(3), 245–257. https://doi.org/10.1016/0016-3287(70)90028-5

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

Beshears, J., & Kosowsky, H. (2020). Nudging: Progress to date and future directions. *Organizational Behavior and Human Decision Processes, 161*, 3–19. https://doi.org/10.1016/j.obhdp.2020.09.001

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as women is to homemaker? Debiasing Word Embeddings. *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain. https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf. Accessed 25 Jan 2023.

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 at https://arxiv.org/abs/2211.13972. Accessed 11 Apr 2023.

Brand, J., Israeli, A., & Ngwe, D. (2023). *Using GPT for market research* (Harvard Business School Marketing Unit Working Paper No. 23–062 and SSRN) at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4395751. Accessed 24 June 2023.

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*. https://doi.org/10.48550/arXiv.1810.03611. Advance online publication.

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*, 980–989. https://doi.org/10.1038/s41562-021-01143-3

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), 2216115120. https://doi.org/10.1073/pnas.2216115120

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

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*. https://doi.org/10.1017/s0140525X22002023. Advance online publication.

Crawford, K. (2021). The hidden costs of AI. *New Scientist, 249*(3327), 46–49. https://doi.org/10.1016/S0262-4079(21)00524-8

De Marcellis-Warn, N., Marty, F., Thelisson, E., Warin, T. (2022). Artificial Intelligence and consumer manipulations: from consumer's counter algorithms to firm's self-regulation tools' *AI and Ethics, 2*, 239–268. https://doi.org/10.1007/s43681-022-00149-5

De Ridder, D., Kroese, F., & van Gestel, L. (2022). Nudgeability: Mapping conditions of susceptibility to nudge influence. *Perspectives on Psychological Science, 17*(2), 346–359. https://doi.org/10.1177/1745691621995183

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica, 90*(1), 81–116. https://doi.org/10.3982/ECTA18709

DeYoung, C. G. (2015). Cybernetic big five theory. *Journal of Research in Personality', 56*, 33–58. https://doi.org/10.1016/j.jrp.2014.07.004

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence, 2*, 423–425. https://doi.org/10.1038/s42256-020-0219-9

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*, 1–16. https://doi.org/10.1016/j.joep.2014.12.003

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), 13347. https://doi.org/10.3390/su132313347

Duckworth, A. L., & Milkman, K. L. (2022). A guide to megastudies. *PNAS Nexus, 1*(5), 1–5. https://doi.org/10.1093/pnasnexus/pgac214

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 at https://arxiv.org/abs/1904.10922. Accessed 15 Sep 2023.

Forrester, J. W. (1971). Counterintuitive behavior of social systems. *Technological Forecasting and Social Change, 3*, 109–140. https://doi.org/10.1016/S0040-1625(71)80001-X

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

Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology, 6*, 361–383. https://doi.org/10.1007/s13164-015-0248-1

Hacker, P. (2021). Manipulation by algorithmics: Exploring the triangle of unfair commercial practice, data protection, and privacy law. *European Law Journal*, 1–34. Advanced online publication. https://doi.org/10.1111/eulj.12389

Hagendorff, T. (2022). Blind spots in AI ethics. *AI and Ethics, 2*, 851–867. https://doi.org/10.1007/s43681-021-00122-8

Hallsworth, M. (2023). A manifesto for applying behavioural science. *Nature Human Behaviour, 7*, 310–323. https://doi.org/10.1038/s41562-023-01555-3

Halpern, D. (2015). *'Inside the Nudge Unit'* W. H. Allen.

Hao, K. (2019, June 6). Training a single AI model can emit as much carbon as five cars in their lifetimes. *MIT Technology Review*. https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/. Accessed 11 Apr 2023.

Hauser, J. R., Liberali, G., & Braun, M. (2009). Website morphing. *Marketing Science, 28*(2), 201–401. https://doi.org/10.1287/mksc.1080.0459

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), 1594–1616. https://doi.org/10.1287/mnsc.2014.1961

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression approach.* Guilford Pres: 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), 2216315120. https://doi.org/10.1073/pnas.2216315120

Helberger, N., Sax, M., Strycharz, J., & Micklitz, H. W. (2022). Choice architectures in the digital economy: Towards a new understanding of digital vulnerability. *Journal of Consumer Policy, 45*, 175–200. https://doi.org/10.1007/s10603-021-09500-5

Hermann, E. (2023) Psychological targeting: nudge or boost to foster midnful and sustainable consumption? *AI and Society, 38*, 961-962. https://doi.org/10.1007/s00146-022-01403-4

Horton, J. J. (2023). *Large language models as simulated economics agents: What can we learn from homo silicus*? ArXiv https://arxiv.org/abs/2301.07543. Accessed 23 June 2023.

Ingrams, A., Kaufmann, W., & Jacobs, D. (2021). In AI we trust? Citizen perceptions of AI in government decision making. *Policy and Internet, 14*(2), 390–409. https://doi.org/10.1002/poi3.276

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), 159–186. https://doi.org/10.1017/bpp.2018.43

Jones, E., & Steinhardt, J. (2022). *Capturing failures of large language models via human cognitive biases.* ArXiv at https://arxiv.org/abs/2202.12299. Accessed 23 June 2023

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgement.* Little & Brown: USA.

Kahneman, D. (2011). *Thinking, fast and slow.* Penguin Books: UK.

Kantorowicz-Reznichenko, E., Kantorowicz, J., & Wells, L. (2022). Can vaccination intentions against COVID-19 be nudged? *Behavioural Public Policy, 11*. Advanced online publication. https://doi.org/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), 145–153. https://doi.org/10.1016/S0140-6736(15)60095-2

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review, 105*(5), 491–495. https://doi.org/10.1257/aer.p20151023

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics, 133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in the age of algorithms. *Journal of Legal Studies, 10*, 113–174. https://doi.org/10.1093/jla/laz001

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*, 2901–2910. https://doi.org/10.1017/pds.2021.551

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), 5802–5805. https://doi.org/10.1073/pnas.1218772110

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), 543–556. https://doi.org/10.1037/a0039210

Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., Herzog, S. (2021) Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. *Humanities and Social Sciences Communications, 8*(1), 117. https://doi.org/10.1057/s41599-021-00787-w

Krpan, D., Makki, F., Saleh, N., Brink, S. I., Klauznicer, H. V. (2021). When behavioural science can make a difference in times of COVID-19. *Behavioural Public Policy, 5*(2):153–179. https://doi.org/10.1017/bpp.2020.48

Krpan, D., & Urbaník, M. (2021). *From libertarian paternalism to liberalism: Behavioural science and policy in an age of new technology*. Advance online publication. https://doi.org/10.1017/bpp.2021.40

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*, 153–179. https://doi.org/10.1017/bpp.2020.48

Krpan, D., Galizzi, M. M., & Dolan, P. (2019). Looking at spillovers in the mirror: Making a case for 'behavioural spillunders'. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01142

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), 721–723. https://doi.org/10.1126/science.1167742

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*, 721–726. https://doi.org/10.1007/s11625-021-00961-8

Lipton, Z. C., & Steinhardt, J. (2018). *Troubling trends in machine learning scholarship.* ArXiv at https://arxiv.org/abs/1807.03341. Accessed 19 Sep 2021.

Ludwig, J. & Mullainathan, S., (2022). *Algorithmic behavioral science: Machine learning as a tool for scientific discovery* (Chicago Booth Working Paper no. 22–15 and SSRN) at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4164272. Accessed 1 Mar 2023.

Ludwig, J., & Mullainathan, S. (2021). Fragile algorithms and fallible decision-makers: Lessons from the justice system. *Journal of Economic Perspectives, 35*(4), 71–96. https://doi.org/10.1257/jep.35.4.71

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), 2200300119. https://doi.org/10.1073/pnas.2200300119

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of ACM Human-Computer Interactions, 3*, 1–32. https://doi.org/10.1145/3359183

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), 12714–12719. https://doi.org/10.1073/pnas.1710966114

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

Meadows, D. (2001). Dancing with systems. *Whole Earth, 106*(3), 58–63.

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). https://doi.org/10.1186/s13012-017-0641-5

Micklitz, H. W., & Pałka, P. (2017). The empire strikes back: Digital control of unfair terms of online services. *Journal of Consumer Policy, 40*, 367–388. https://doi.org/10.1007/s10603-017-9353-0

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* ArXiv at https://arxiv.org/pdf/1301.3781.pdf. Accessed 4 Jul 2021.

Mills, S. (2022). Personalized nudging. *Behavioural Public Policy, 6*(1), 150–159. https://doi.org/10.1017/bpp.2020.7

Mills, S. (2022). Finding the 'nudge' in hypernudge. *Technology in Society, 71*, 102117. https://doi.org/10.1016/j.techsoc.2022.102117

Mills, S., & Sætra, H. S. (2022). *The autonomous choice architect*. Advance online publication. https://doi.org/10.1007/s00146-022-01486-z

Mills, S., & Whittle, R. (2023). *Seeing the nudge from the trees: The 4S framework for evaluating nudges*. Advance online publication. https://doi.org/10.1111/padm.12941

Morozovaite, V. (2021). Two sides of the digital advertising coin: Putting hypernudging into perspective. *Market and Competition Law Review, 5*(2), 105–145. https://doi.org/10.34632/mclawreview.2021.10307

Mullainathan, S., & Obermeyer, Z. (2022). Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics, 137*(2), 679–727. https://doi.org/10.1093/qje/qjab046

Ng, C. F. (2016). Behavioral mapping and tracking. In Gifford, R. (Eds.) *Research methods for environmental psychology*. https://doi.org/10.1002/9781119162124.ch3

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*, 3901. https://doi.org/10.1038/s41467-020-17614-6

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 at https://arxiv.org/pdf/2304.03442.pdf. Accessed 20 Apr 2023.

Pedersen, T., & Johansen, C. (2020). Behavioural artificial intelligence: An agenda for systematic empirical studies of artificial inference. *AI and Society, 35*(3), 519–532. https://doi.org/10.1007/s00146-019-00928-5

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*, 106347. https://doi.org/10.1016/j.chb.2020.106347

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., … Wellman, M. (2019). Machine behaviour. *Nature, 568*, 477–486. https://doi.org/10.1038/s41586-019-1138-y

Rauthmann, J. F. (2020). A (more) behavioural science of personality in the age of multi-modal sensing, big data, machine learning, and artificial intelligence. *European Journal of Personality, 34*(5), 593–598. https://doi.org/10.1002/per.2310

Rebonato, R. (2014). A critical assessment of libertarian paternalsim. *Journal of Consumer Policy, 37*, 357–396. https://doi.org/10.1007/s10603-014-9265-1

Russell, S. J. (2019). *Human compatible: AI and the problem of control.* Penguin Books: UK.

Sætra, H. S. (2020). Privacy as an aggregate public good. *Technology in Society, 63*, 101422. https://doi.org/10.1016/j.techsoc.2020.101422

Saheb, T. (2022). *Ethically contentious aspects of artificial intelligence surveillance: A social science perspective*. Advance online publication. https://doi.org/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*. https://doi.org/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), 672–690. https://doi.org/10.1177/1745691619848688

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), 144–167. https://doi.org/10.1017/bpp.2018.17

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*. Advance online publication. https://doi.org/10.1017/bpp.2022.40

Schmidt, R., & Stenger, K. (2021). *Behavioral brittleness: The case for strategic behavioral public policy*. Advance online publication. https://doi.org/10.1017/bpp.2021.16

Simon, H. A. (1981). *The sciences of the artificial* (2nd ed.). MIT Press.

Smith, J., & de Villiers-Botha, T. (2021). *Hey, Google, leave those kids alone: Against hypernudging children in the age of big data*. Advance online publication. https://doi.org/10.1007/s00146-021-01314-w

Spencer, N. (2018). Complexity as an opportunity and challenge for behavioural public policy. *Behavioural Public Policy, 2*(2), 227–234. https://doi.org/10.1017/bpp.2018.20

Strubell, E., Verga, P., Andor, D., Weiss, D., & McCallum, A. (2018). *Linguistically-informed self-attention for semantic role labelling*. ArXiv at https://arxiv.org/abs/1804.08199. Accessed 11 Apr 2023.

Sunstein, C. R. (2022). The distributional effects of nudges. *Nature Human Behaviour, 6*, 9–10. https://doi.org/10.1038/s41562-021-01236-z

Sunstein, C. R. (2022). Governing by algorithm? No noise and (potentially) less bias. *Duke Law Journal, 71*(6), 1175–1205.

Sunstein, C. R. (2023). The use of algorithms in society. *The Review of Austrian Economics*. https://doi.org/10.1007/s11138-023-00625-z. Advance online publication.

Sunstein, C. R. (2012). *Impersonal default rules vs. active choices vs. personalized default rules: A triptych.* SSRN at https://dash.harvard.edu/bitstream/handle/1/9876090/decidingbydefault11_5.pdf?sequence=1. Accessed 11 Apr 2023.

Sunstein, C. R. (2015). *The ethics of influence.* Cambridge University Press: USA.

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), 2200732119. https://doi.org/10.1073/pnas.2200732119

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *American Economic Review, 93*, 175–179. https://doi.org/10.1257/000282803321947001

Thaler, R. H., Sunstein, C. R. (2008). 'Nudge: Improving Decisions about Health, Wealth, and Happiness' Penguin Books

Thorun, C., & Diels, J. (2020). Consumer protection technologies: An investigation into the potentials of new digital technologies for consumer policy. *Journal of Consumer Policy, 43*, 177–191. https://doi.org/10.1007/s10603-019-09411-6

Thunström, L. (2019). Welfare effects of nudges: The emotional tax of calorie menu labeling. *Judgment and Decision Making, 14*(1), 11–25. https://doi.org/10.1017/S1930297500002874

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*, 267–282. https://doi.org/10.1016/j.jebo.2018.07.005

Tierney, W., Hardy, J. H., Ebersole, C. R., Leavitt, K., Viganola, D., Clemente, E. G., Gordon, M., Dreber, A., Johannesson, M., Pfeiffer, T., Collaboration, H. D. F., & Uhlmann, E. L. (2020). Creative destruction in science. *Organizational Behavior and Human Decision Processes, 161*, 291–309. https://doi.org/10.1016/j.obhdp.2020.07.002

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

UK Centre for Data Ethics and Innovation (2020). Review into bias in algorithmic decision-making. UK Government. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957259/Review_into_bias_in_algorithmic_decision-making.pdf. Accessed 21 Apr 2023.

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

Wilke, A., & Mata, R. (2012). Cognitive bias. In *Encyclopaedia of human behaviour* (2nd ed.) https://doi.org/10.1016/B978-0-12-375000-6.00094

Wolfram, S. (2023). What is ChatGPT doing… and why does it work? https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/. Accessed 17 Feb 2023.

Yeung, K. (2017). Hypernudge: Big data as a mode of regulation by design. *Information Communication and Society, 1*, 118–136. https://doi.org/10.1080/1369118X.2016.1186713