Humanities & Social Sciences Communications

COMMENT

https://doi.org/10.1057/s41599-024-03428-0 OPEN

Check for updates

Choice engines and paternalistic AI

Cass R. Sunstein [1]✉

Many consumers suffer from inadequate information and behavioral biases, which can produce internalities, understood as costs that people impose on their future selves. In these circumstances, "Choice Engines," powered by Artificial Intelligence (AI), might produce significant savings in terms of money, health, safety, or time. Different consumers care about different things, of course, which is a reason to insist on a high degree of freedom of choice, and a high degree of personalization. Nonetheless, it is important to emphasize that Choice Engines and AI might be enlisted by self-interested actors, who might exploit inadequate information or behavioral biases, and thus reduce consumer welfare. It is also important to emphasize that Choice Engines and AI might show behavioral biases, perhaps the same ones that human beings are known to show, perhaps others that have not been named yet, or perhaps new ones, not shown by human beings, that cannot be anticipated.

[1] Harvard University, Cambridge, USA. ✉email: csunstei@law.harvard.edu

Heterogeneity and personalization

Can artificial intelligence increase social welfare by improving people's choices? Can it address the problem of heterogeneity? Might self-interested designers or users exploit behavioral biases, increase manipulation, and thus reduce social welfare? The answer to all of these questions is "yes"—which raises serious issues for regulators, and serious empirical challenges for researchers.

Consider three sets of findings:

1. On average, people appear to benefit from home energy reports; they save money, on average, and they are willing to pay, on average, a positive amount to receive such reports. But some people are willing to pay far more than others (Allcott and Kessler, 2019). In fact, some people are willing to pay *not* to receive home energy reports. They believe that they would be better off without them. While home energy reports are designed to save consumers money and reduce externalities, sending such reports to some people seems to have costs in excess of benefits. More targeted use of home energy reports could produce significant welfare gains (Allcott and Kessler, 2019).
2. On average, graphic warning labels on sugary drinks affect consumer behavior, and in what seems to be the right way; such labels reduce demand for such drinks. At the same time, labels on sugary drinks have greater effects on some consumers than on others (Allcott et al., 2022). Disturbingly, such labels can lead people who do not have self-control problems to consume less in the way of sugary drinks, while having a significantly smaller effect on people who do have self-control problems. In addition, many people do not like seeing graphic warning labels. The average person in a large sample reported being willing to pay about $1 to *avoid* seeing the graphic warning labels (Allcott et al., 2022). It is likely that such labels help some and hurting others. It is possible that such labels on balance cause harm.
3. There is evidence that calorie labels have welfare benefits (Thunström, 2019). At the same time, they seem to have a greater effect on people who lack self-control problems than on people who suffer from such problems. It is possible that in some populations, calorie labels affect people who do not need help, and have little or no effect on people who do need help, except to make them feel sad and ashamed.

From these sets of findings, we can draw three simple conclusions. *First*, interventions may have either positive or negative *hedonic* effects. People might like seeing labels, or they might dislike seeing labels. *Second*, interventions might well have different effects on different populations. Under favorable conditions, they might have large positive effects on a group that needs help, and small or no effects on a group that does not need help. Under unfavorable conditions, they might have small or no effects on a group that needs help, and large effects on a group that does not need help. Large effects on a group that does not need help may not improve that group's welfare. For example, people who have no need to change their spending patterns, or their diets, might end up doing so. *Third*, and consistent with the second conclusion, an understanding of the average treatment effect does not tell us what we need to know (Allcott et al., 2022). Personalization can produce significant welfare gains.

These points about labels can be made about a wide range of interventions. They hold for automatic enrollment: It is possible that automatic enrollment in some plans will have no effects on people who benefit from enrollment (because they would enroll in any case) while harming people who do not benefit from enrollment (because some or many who lose do not opt-out, perhaps because of inertia). They hold for taxes: It is possible that taxes will have little or no effect on the people they are particularly intended to help, while having a significant adverse effect on people they are not (particularly) intended to help. (Consider soda taxes.) They hold for mandates and bans: A ban on some activity or product might, on balance, hurt people who greatly benefit from that activity or product, while helping people who lose only modestly from it. (Consider bans on the purchase of incandescent lightbulbs, or a prohibition on gasoline-powered cars, and put externalities to one side.) In all of these cases, more targeted action and greater personalization would be far better than "mass" action.

Choice engines can increase welfare

To understand the promise of AI, note that for retirement plans, many employers use something like a Choice Engine.[1] They know a few things about their employees (and possibly more than a few). On the basis of what they know, they automatically enroll their employees in a specific plan. The plan is frequently a diversified, passively managed index fund. Employees can opt-out and choose a different plan if they like. Alternatively, employers might offer employees a specified set of options, with the understanding that all of them are suitable, or suitable enough. (Options that are not suitable are not included.) They might provide employees with simple information by which to choose among them. The options might be identified or rethought with the assistance of artificial intelligence (AI) or some kind of algorithm (see Fidelity, n.d.).

Here is one reasonable approach: Automatically enroll employees in a plan that is most likely to improve their well-being, given everything relevant that is known about them.[2] Identification of that plan might prove daunting, but a large number of plans can at least be ruled out (Ayres and Curtis, 2023). Note that if the focus is on improving employee well-being, we are not necessarily speaking of revealed preferences.

For retirement savings, we can easily imagine many different kinds of Choice Engines (see Ayres and Curtis, 2023). Some of them might be mischievous; some of them might be fiendish; some of them might be random; some of them might be coarse or clueless; some of them might show behavioral or other biases of their own; some of them might be self-serving.[3] For example, people might be automatically enrolled in plans with high fees. They might be automatically enrolled in plans that are not diversified. They might be automatically enrolled in money market accounts. They might be automatically enrolled in dominated plans (Ayres and Curtis, 2023). They might be automatically enrolled in plans that are especially ill-suited to their situations. They might be given a large number of options and asked to choose among them, with little relevant information, or with information that leads them to make poor choices.

I mean to use this example to offer a general point: In principle, Choice Engines, powered by AI, might work to overcome an absence of information and behavioral biases (Hasan et al., 2023),[4] and they might also be highly personalized. For retirement plans, Choice Engines may or may not be paternalistic. If they are not paternalistic, it might be because they simply provide a menu of options, given what they know about relevant choosers (see Purina, n.d.). If they are paternalistic, they might be mildly paternalistic, moderately paternalistic, or highly paternalistic. A moderately paternalistic Choice Engine might impose nontrivial barriers on those who seek certain kinds of plans (such as those with high fees). The barriers might take the form of information provision, "are you sure you want to?" queries, and requirements of multiple clicks. We might think of a moderately paternalistic Choice Engine as offering "light patterns," as contrasted with

"dark patterns" (Luguri and Strahilevitz, 2021). A highly paternalistic Choice Engine might forbid employees from selecting any plan other than the one that it deems in the interest of employees or might make it exceedingly difficult for employees to do that.

Choice Engines of this kind might be used for any number of choices, including (to take some random examples) choices of dogs, laptops, mystery novels, cellphones, shavers, shoes, tennis racquets, and ties (see Purina, n.d.). Choice Engines may or may not use AI, and if they do, they can use AI of different kinds. Consider this question: What kind of car would you like to buy? Would you like to buy a fuel-efficient car that would cost you $800 more upfront than the alternative but that would save you $8000 over the next ten years? Would you like to buy an energy-efficient refrigerator that would cost you $X today, but save you ten times $X over the next ten years? What characteristics of a car or a refrigerator matter most to you? Do you need a large car? Do you like hybrids? Are you excited about electric cars, or not so much?

A great deal of work finds that consumers suffer from "present bias" (Schleich et al., 2019; Werthschulte and Löschel, 2021; Kuchler and Pagel, 2018; O'Donoghue and Rabin, 2015; Benhabib et al., 2010; Wang and Sloan, 2018).[5] Current costs and benefits loom large; future costs and benefits do not. For many of us, the short-term is what matters most, and the long-term is a foreign country. The future is Laterland, a country that we are not sure that we will ever visit. This is so with respect to choices that involve money, health, safety, and more.[6]

Artificial intelligence (AI) need not suffer from present bias.[7] Imagine that you are able and willing to consult AI to ask it what kind of car you should buy. Imagine too that you discover that you are, or might be, present-biased, in the sense that you prefer a car that is not (according to AI) the one that you should get. What then? We could easily imagine Choice Engines for motor vehicle purchases in which different consumers provide relevant information about their practices, their preferences, and their values, and in which the relevant Choice Engine immediately provides a set of options—say, Good, Better, and Best. Something like this could happen in minutes or even seconds, perhaps a second or two. If there are three options—Good, Better, and Best—verbal descriptions might explain the ranking. Or a Choice Engine might simply say: Best For You. It might do so while allowing you to see other options if you indicate that you wish to do so. It may or may not be paternalistic, or come with guardrails designed to protect consumers against serious mistakes (Ayres and Curtis, 2023).

Internalities, externalities, and personalization

Attempting to respond to the kinds of findings with which I began, those who design Choice Engines might well focus solely on particular consumers and what best fits their particular situations. They might ask, for example, about what particular consumers like most in cars, and they might take into account the full range of economic costs, including the costs of operating a vehicle over time. If so, choice engines would be highly personalized.

They might also have a paternalistic feature insofar as they suggest that Car A is "best" for a particular consumer, even if that consumer would not give serious consideration to Car A. A Choice Engine would attempt to overcome both informational deficits and behavioral biases on the part of those who use them. Freedom of choice would be preserved, in recognition of the diversity of individual tastes, including preferences and values.

Present bias is, of course, just one reason that consumers might not make the right decisions, where "right" is understood by reference to their own welfare. Consumers might also suffer from a simple absence of information, from status quo bias, from limited attention, or from unrealistic optimism. If people are making their own lives worse for any of these reasons, Choice Engines might help. They might be paternalistic insofar as they respond to behavioral biases on the part of choosers, perhaps by offering recommendations or defaults, perhaps by imposing various barriers to choices that, according to the relevant Choice Engine, would not be in the interest of choosers.

Alternatively, Choice Engines might take account of externalities. Focusing on greenhouse gas emissions, for example, they might use the social cost of carbon to inform choices. Suppose, for simplicity, that it is $100. Choice Engines might select Good, Better, and Best, incorporating that number. A Choice Engine that includes externalities might do so by default, or it might do so if and only if choosers explicitly request it to do so.

Choice Engines might be designed in different ways. They might allow consumers to say what they care about—including or excluding externalities, for example. They might be designed so as to include externalities, but to be transparent about their role, allowing consumers to see Good, Better, and Best with and without externalities. They might be designed so as to allow a great deal of transparency with respect to when costs would be incurred. If, for example, a car would cost significantly more upfront, but significantly less over a period of five years, a Choice Engine could reveal that fact.

We could imagine a Keep It Simple version of a Choice Engine, offering only a little information and a few options to consumers. We could imagine a Tell Me Everything version of a Choice Engine, living up to its name. Consumers might be asked to choose what kind of Choice Engine they want. Alternatively, they might be defaulted to Keep It Simple or Tell Me Everything, depending on what AI thinks they would choose, if they were to make an informed choice, free from behavioral biases. Personalization on this count would have major advantages.

Dangers and risks

To be sure, there are dangers and risks. Consider three points:

1. Those who design Choice Engines, or anything like them, might be self-interested or malevolent. Rather than correcting an absence of information or behavioral biases, they might *exploit* them. Algorithms and AI threaten to do exactly that, in a way that signals the presence of manipulation (Bar-Gill et al., 2023). Indeed, AI could turn out to be highly manipulative, thus harming consumers (Sunstein, 2022). This is a potentially serious threat, not least when personalization is combined with manipulation.
2. Choice Engines might turn out to be coarse; they might replicate some of the problems of "mass" interventions. They may or may not be highly personalized. If they use a few simple cues, such as age and income, they might not have the expected or hoped-for welfare benefits. Algorithms or AI might turn out to be insufficiently informed about the tastes and values of particular choosers (Rizzo and Whitman, 2020).
3. Whether paternalistic or not, AI might turn out to suffer from its own behavioral biases. There is evidence that LLMs show some of the biases that human beings do (Chen et al., 2023). It is possible that AI will show biases that human beings show that have not even been named yet. It is also possible that AI will show biases of its own.

For these reasons, the same kinds of guardrails that have been suggested for retirement plans might be applied to Choice Engines of multiple kinds, including those involving motor vehicles and appliances (Ayres and Curtis, 2023). Restrictions on

the equivalent of "dominated options," for example, might be imposed by law, so long as it is clear what is dominated (Bhargava et al. 2017). Restrictions on shrouded attributes, including hidden fees, might be similarly justified (Ayres and Curtis, 2023). Choice Engines powered by AI have considerable potential to improve consumer welfare and also to reduce externalities, but without regulation, we have reason to question whether they will always or generally do that (Akerlof and Shiller, 2015). Those who design Choice Engines may or may not count as fiduciaries,[8] but at a minimum, it makes sense to scrutinize all forms of choice architecture for deception and manipulation, broadly understood.

Received: 8 April 2024; Accepted: 1 July 2024;
Published online: 06 July 2024

Notes

1 The term is not in general use, but something like it can be found in various places, with variations (see Yeomans et al., 2019; Champniss, 2018; compare to Whirlpool, n.d.).
2 This is consistent with Ayres and Curtis (2023).
3 An episode of Black Mirror could easily be based on such scenarios.
4 For a disturbing set of findings, see Chen et al. (2023).
5 Importantly, Wang and Sloan (2018) find strong evidence of present bias in connection with health-related decisions.
6 There are plausible evolutionary explanations for present bias. If you are running from a tiger, you ought not to spend much time thinking about your retirement savings. But under modern circumstances, present bias can get you into a great deal of trouble.
7 It might (Chen et al., 2023).
8 For more information about how the law views those who design Choice Engines, consider Hughes vs. Northwestern University, 142S. Ct. 737 (2022) and Hughes vs. Northwestern University, 63F.4th 615 (7th Cir. 2023).

References

Akerlof GA, Shiller RJ (2015) Phishing for phools: the economics of manipulation & deception. Princeton University Press, Princeton
Allcott H, Cohen D, Morrison W, Taubinsky D (2022) When do "nudges" increase social welfare? NBER working paper no. 30740. Available via NBER. https://www.nber.org/system/files/working_papers/w30740/w30740.pdf. Accessed 25 Mar 2024
Allcott H, Kessler JB (2019) The welfare effects of nudges. Am Econ J: Appl Econ 11(1):236–276. https://doi.org/10.1257/app.20170328
Ayres I, Curtis Q (2023) Retirement guardrails. Cambridge University Press, Cambridge
Bar-Gill O, Sunstein CR, Talgam-Cohen I (2023) Algorithmic harm in consumer markets. Available via SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4321763. Accessed 25 Mar 2024
Benhabib J, Bisin A, Schotter A (2010) Present-bias, quasi-hyperbolic discounting, and fixed costs. Games Econ Behav 69(2):205–223. https://doi.org/10.1016/j.geb.2009.11.003
Bhargava S, Loewenstein G, Sydnor J (2017) Choose to lose: health plan choices from a menu with dominated option. Q J Econ 132(3):1319–1372. https://doi.org/10.1093/qje/qjx011
Champniss G (2018) The rise of the choice engine. https://www.enervee.com/blog/the-rise-of-the-choice-engine. Accessed 25 Mar 2024
Chen Y, Andiappan M, Jenkin T, Ovchinnikov A (2023) A manager and an AI walk into a bar: does ChatGPT make biased decisions like we do? Available via SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4380365. Accessed 25 Mar 2024
Fidelity (n.d.) Fidelity Go. https://www.fidelity.com/managed-accounts/fidelity-go/overview. Accessed 25 Mar 2024
Hasan Z, Vaz D, Athota VS, Maturin Désiré SS, Pereira V (2023) Can artificial intelligence (AI) manage behavioural biases among financial planners? J Glob Inf Manag 31(2):1–18. https://doi.org/10.4018/JGIM.321728
Kuchler T, Pagel M (2018) Sticking to your plan: the role of present bias for credit card paydown. NBER Working Paper No. 24881. Available via NBER. https://www.nber.org/system/files/working_papers/w24881/w24881.pdf. Accessed 25 Mar 2024
Luguri J, Strahilevitz LJ (2021) Shining a light on dark patterns. J Leg Anal 13(1):43–109. https://doi.org/10.1093/jla/laaa006
O'Donoghue T, Rabin M (2015) Present bias: lessons learned and to be learned. Am Econ Rev: Pap Proc 105(5):273–279. https://doi.org/10.1257/aer.p20151085
Purina (n.d.) Welcome to the Purina dog breed selector. https://www.purina.co.uk/find-a-pet/dog-breeds/breed-selector. Accessed 25 Mar 2024
Rizzo MJ, Whitman G (2020) Escaping paternalism: rationality, behavioural economics, and public policy. Cambridge University Press, Cambridge
Schleich J, Gassmann X, Meissner T, Faure C (2019) A large-scale test of the effects of time discounting, risk aversion, loss aversion, and present bias on household adoption of energy-efficient technologies. Energy Econ 80:377–393. https://doi.org/10.1016/j.eneco.2018.12.018
Sunstein CR (2022) Manipulation as theft. J Eur Public Policy 29(12):1959–1969. https://doi.org/10.1080/13501763.2022.2135757
Thunström L (2019) Welfare effects of nudges: the emotional tax of calorie menu labeling. Judgm Decis Mak 14(1):11–25. https://doi.org/10.1017/S1930297500002874
Wang Y, Sloan FA (2018) Present bias and health. J Risk Uncertain 57(2):177–198. https://doi.org/10.1007/s11166-018-9289-z
Werthschulte M, Löschel A (2021) On the role of present bias and biased price beliefs in household energy consumption. J Env Econ Manag 109. https://doi.org/10.1016/j.jeem.2021.102500
Whirlpool (n.d.) Buying a refrigerator guide: how to choose a new fridge in 2024. https://www.whirlpool.com/blog/kitchen/buying-guide-refrigerator.html. Accessed 25 Mar 2024
Yeomans M, Shah A, Mullainathan S, Kleinberg J (2019) Making sense of recommendations. J Behav Decis Mak 32(4):403–414. https://doi.org/10.1002/bdm.2118

Acknowledgements

Some parts of this essay draw on previous work on fuel economy regulation. I am grateful to the Harvard Law School's Program on Behavioral Economics and Public Policy and the Harvard Law School's Initiative on Artificial Intelligence and Law for valuable support.

Additional information

Correspondence and requests for materials should be addressed to Cass R. Sunstein.