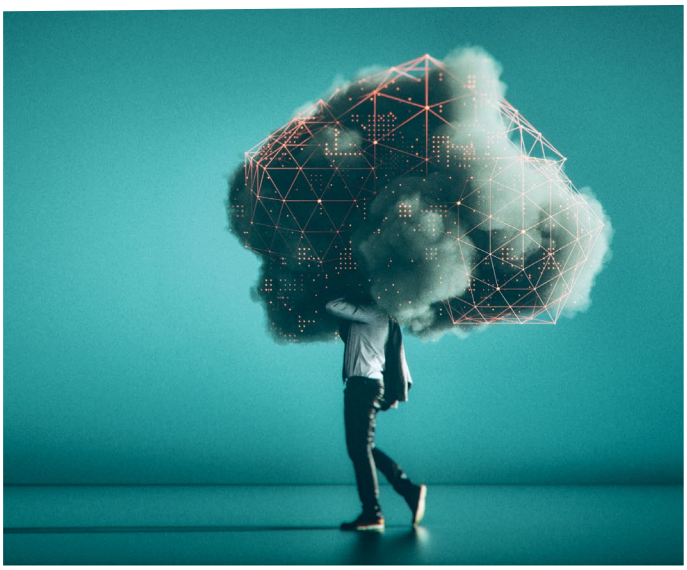
https://doi.org/10.1038/s41562-023-01724-4

Human bias in algorithm design

Carey K. Morewedge, Sendhil Mullainathan, Haaya F. Naushan, Cass R. Sunstein, Jon Kleinberg, Manish Raghavan & Jens O. Ludwig

Check for updates

Algorithms are designed to learn user preferences by observing user behaviour. This causes algorithms to fail to reflect user preferences when psychological biases affect user decision making. For algorithms to enhance social welfare, algorithm design needs to be psychologically informed.



Many people believe that algorithms are failing to live up to their promise to reflect user preferences and improve social welfare[1–4]. The problem is not technological. Modern algorithms are sophisticated and accurate. Training algorithms on unrepresentative samples contributes to the problem, but failures happen even when algorithms are trained on the population. Nor is the problem caused only by the profit motive. For-profit firms design algorithms at a cost to users, but even non-profit organizations and governments fall short[5].

All algorithms are built on a psychological model of what the user is doing. The fundamental constraint on this model is the narrowness of the measurable variables for algorithms to predict. We suggest that algorithms fail to reflect user preferences and enhance their welfare because algorithms rely on revealed preferences to make predictions. Designers build algorithms with the erroneous assumption that user behaviour (revealed preferences) tells us (1) what users rationally prefer (normative preferences) and (2) what will enhance user welfare. Reliance on this 95-year-old economic model, rather than the more realistic assumption that users exhibit bounded rationality, leads designers to train algorithms on user behaviour. Revealed preferences can identify unknown preferences, but revealed preferences are an incomplete — and at times misleading — measure of the normative preferences and values of users[6]. It is ironic that modern algorithms are built on an outmoded and indefensible commitment to revealed preferences.

What algorithms learn from behaviour

Over the past five decades, psychologists and behavioural economists have documented many anomalies in human decision making; systematic deviations from assumptions of revealed preferences. When user behaviour exhibits these biases, algorithms will wrongly conflate revealed and normative preferences. Below, we consider three examples.

Fast thinking. Users often lack the knowledge, time, capacity or motivation to decide rationally. When these constraints bound rationality, users often rely on associative intuitions and habits and are influenced by contextual factors such as choice defaults[7]. Relying on these decision strategies is typically adaptive but can create systematic biases in user decision making[2]. Some of these biases are cognitive, and some involve discrimination[3]. Algorithms trained on user behaviour, such as hiring decisions, will reflect human biases and structural inequities in systems that people do not endorse and may be unaware of. Indeed, when Amazon trained a hiring algorithm on its past hiring decisions, the algorithm revealed gender bias that had escaped notice when those human hiring decisions were unaggregated[8]. Algorithms trained on habitual behaviour may learn preferences that people no longer endorse or never endorsed. For example, most smokers want to quit. When no control group exists, algorithms are often blind to contextual influences that misalign preferences and behaviour. If retirement savings are larger when deductions are automatically withheld from future raises and smaller when deductions reduce present income, an algorithm trained in the former context would infer that users prefer to save more. An algorithm trained in the latter context would infer that users prefer to save little.

Wants, not shoulds. People hold conflicting desires and motives — to live for the present or the future, take for themselves or share, or to expand or exploit their knowledge. Algorithms observe and base recommendations on user preferences when decisions are made, and not on user preferences in foresight or hindsight. Consequently, algorithms learn the conflict resolutions that are most immediately rewarding ('wants') rather than the most rewarding long-term resolutions ('shoulds'). Netflix users tend to fill their watch lists with highbrow films — an explicit expression of their 'shoulds' — but end up streaming the lowbrow shows and movies that its algorithm recommends[9]. Users learn from some of these recommendations (such as rational epistemic actions) and enjoy many of them in the moment, and companies that cater to 'wants' benefit from high consumer demand (for example, news and social media platforms serving clickbait), but platforms can 'get you' and simultaneously 'be ruining your life'. User engagement with social media content that evokes moral outrage offers an example of

CREDIT: GREMLIN/E+/GETTY

pernicious long-term effects on user well-being and social harmony. A more balanced approach would better serve the interests of users and firms.

Social norms and the status quo. Users rely on algorithmic curation to find and discover options in the large catalogues that platforms offer. Recommendations and lists, such as bestsellers or critic's choices, guide option search and discovery. Users are recommended options on the basis of their past choices (content-based filtering) or options chosen by users whose choices are similar to their choices (collaborative filtering). Recommendation systems require many observations to make accurate predictions. Limited data constrain their ability to recommend and infer revealed preferences for niche options, new options and options that have yet to diffuse through markets (for example, alternative-fuel vehicles). Consequently, recommendations and lists prioritize popular and existing options. Without considerable engineering, Harry Potter would be recommended everywhere whether it is relevant (for example, to users watching *The Lord of the Rings*) or irrelevant (for example, to users reading *Mastering the Art of French Cooking*). It is contested whether recommendations systems are creating a monoculture, echo chambers or 'filter bubbles' in what users consume and firms produce. Evidence from narrower domains shows that recommendation systems change user preferences and reduce the diversity of what consumers see and buy, which increases the market share of popular options[3,4].

A central problem is that algorithms are trained to learn from revealed preferences. User behaviour provides the large volume of data needed to train algorithms and it soothes economists who have, since Pareto, been sceptical of any measure but observed behaviour[6]. But revealed preferences have a deeply Skinnerian feel and are subject to the same objections as behaviourism: whether revealed preferences reflect normative preferences depends on (among other things) other concurrently active preferences and mental states of a user[10]. Users might value learning about the economy and climate but feel that reading entertainment and sports news is all they can handle after work. The friends who users talk with may have different political affiliations or may prefer talking about entertainment and sports. Users might read news about entertainment or sports because they feel less confident in their understanding of the economy or climate science. Of course, it is possible that user behaviour reflects normative preferences, but it is also possible that it does not. Reliance on revealed preferences might disserve users and fail to promote their welfare.

Even advanced large language models such as ChatGPT run into this problem because they learn from biased user behaviour[11]. Large language models are prone to associative hallucinations, for instance, and hallucinate inaccurate answers that occur frequently in their training data[12].

Ways forward

There are better ways to design algorithms to reflect normative preferences. Algorithmic audits can reveal when algorithmic predictions are likely to reflect normative preferences and when they are likely to reflect biased associations, intuitions and context (and how to annotate user behaviour accordingly). For instance, an algorithmic audit revealed that recommendations made by a Facebook algorithm trained on more-intuitive user behaviour (newsfeed content consumed) exhibited more out-group bias than an algorithm trained on more-deliberative user behaviour (people who they friended)[3]. External audits can be conducted with user samples, publicly available API (application programming interface) data, and bots[4].

Experiments are the gold standard for testing whether revealed preferences reflect normative preferences or habits and self-control failures. Researchers without the cooperation of firms can conduct user intervention studies. Facebook users who were paid to deactivate their account for four weeks reduced their usage after this period by 22% and reported an increase in their subjective well-being[1], which suggests that part of their usage was habitual. Habitual behaviours and learning can also be uncovered by comparing new and experienced user behaviour. Cleaner experiments that better preserve user privacy (A–B tests) require collaboration with firms[4]. Algorithms could be tuned from reflecting 'wants' toward 'shoulds' by expanding the time horizon of observed behaviour. Algorithms could be trained on simulated and real users who exhibit better decision making or desired states, such as safer drivers or happier users. Algorithms can increase option diversity by increasing options and recommendation randomness or balancing the weights given to in-network and out-of-network recommendations (for example, the behaviour of similar and dissimilar users). Algorithms can handle data drift when behaviour changes – more difficult is concept drift. When revealed preferences and normative preferences change, retraining or a new model is necessary.

Preference elicitation methods, ranging from user surveys to conjoint analyses, can identify normative preferences and new objectives to optimize; however, fully replacing user behaviour with stated preferences is not a solution. Decades of psychological research shows that people often cannot explain their preferences. Similar to perceptual decisions ('this salad is green'), people have access to the output of intuitive decisions ('those nachos are better') but often lack access to the processes by which they made intuitive decisions[7]. Expanding the basket of objectives that algorithms predict beyond revealed preferences and simple feedback such as likes and shares to include stated preferences (self-reports) and correlates of psychological states such as content sentiment, search queries or user interaction speeds could make recommendation systems better reflect normative preferences and enhance user well-being. Comparisons between stated and revealed preferences can identify when revealed and normative preferences overlap and diverge. Models that mix stated and revealed preferences or structural models could help to design more psychologically informed algorithms[6,13].

It is time to invest in the behavioural science of algorithm design. A paradigmatic advance in algorithms is unlikely to come from having more data. Algorithm design should move beyond revealed preferences for algorithms to better reflect normative preferences and deliver on their potential to improve social welfare.

Carey K. Morewedge [1] ✉, Sendhil Mullainathan[2], Haaya F. Naushan [3], Cass R. Sunstein [4], Jon Kleinberg[5], Manish Raghavan[6] & Jens O. Ludwig [7]

[1]Questrom School of Business, Boston University, Boston, MA, USA. [2]Chicago Booth School of Business, University of Chicago, Chicago, IL, USA. [3]The World Bank, Toronto, Ontario, Canada. [4]Harvard Law School, Harvard University, Cambridge, MA, USA. [5]Department of Computer Science, Cornell University, Ithaca, NY, USA. [6]MIT Sloan School of Management, MIT, Cambridge, MA, USA. [7]Harris School of Public Policy, University of Chicago, Chicago, IL, USA.

✉e-mail: morewedg@bu.edu

Published online: 20 November 2023

References

1. Allcott, H., Braghieri, L., Eichmeyer, S. & Gentzkow, M. *Am. Econ. Rev.* **110**, 629–676 (2020).
2. Agan, A. Y., Davenport, D., Ludwig, J. & Mullainathan, S. *Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias (No. w30981)* (National Bureau of Economic Research, 2023).
3. Lee, D. & Hosanagar, K. *Inf. Syst. Res.* **30**, 239–259 (2019).
4. GPAI. *Transparency Mechanisms for Social Media Recommender Algorithms: From Proposals to Action. Tracking GPAI's Proposed Fact Finding Study in This Year's Regulatory Discussions* (Global Partnership on AI, 2022).
5. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. *Science* **366**, 447–453 (2019).
6. Beshears, J., Choi, J. J., Laibson, D. & Madrian, B. C. *J. Public Econ.* **92**, 1787–1794 (2008).
7. Morewedge, C. K. & Kahneman, D. *Trends Cogn. Sci.* **14**, 435–440 (2010).
8. Logg, J. M. Using algorithms to understand the biases in your organization. *Harv. Bus. Rev.*, https://hbr.org/2019/08/using-algorithms-to-understand-the-biases-in-your-organization (9 August 2019).
9. Milkman, K. L., Rogers, T. & Bazerman, M. H. *Manage. Sci.* **55**, 1047–1059 (2009).
10. Block, N. *Philos. Rev.* **90**, 5–43 (1981).
11. Ray, P. P. *Internet Things Cyber-Phys. Syst.* **3**, 121–154 (2023).
12. McKenna, N. et al. Preprint at *arXiv*, https://doi.org/10.48550/arXiv.2305.14552 (2023).
13. Kleinberg, J., Ludwig, J., Raghavan, M. & Mullainathan, S. *Perspect. Psychol. Sci.* (in the press).

Competing interests

The authors declare no competing interests.