Revisiting Asch: Examining Conformity to Algorithms and Algorithm Aversion

Benjamin Lira[1]

[1] University of Pennsylvania

Author Note

Social Psychology - PSYC0600-3.

Benjamin Lira https://orcid.org/0000-0001-5328-0657

Correspondence concerning this article should be addressed to Benjamin Lira, 425 S University Ave, Philadelphia, PA 19104. E-mail: blira@upenn.edu

Abstract

Algorithm aversion refers to the phenomenon where people prefer human decisions as opposed to algorithmic decisions or recommendations. For example, people might distrust an algorithm's prediction of a house price, and prefer to trust a real-estate agent's judgment. The literature on conformity suggests that people tend to agree when multiple others provide an answer, even when it disagrees with what they would have considered correct absent social influence. In this investigation, we test the possibility of conformity to a large number of agreeing algorithms as a potential remedy to algorithm aversion. Subjects are presented with descriptions and pictures of houses and are asked to put a price to each house. In one condition, they provide the price of the home, in others they do so after seeing recommendations from one algorithm (e.g., Zillow's home price algorithm), or more than one agreeing algorithms (e.g., Zillow, Google, and Facebook's algorithms). Unbeknownst to the subject, the "algorithmic predictions" are actually manipulated to be either too high or too low compared to the real house value. We could vary the number of algorithms (0, 1, 3, 5), their level of agreement (SD = 100, 1,000, or 10,000), and their distance from the real house price (10% over/under, double/half real price). Results would provide insight on how people interact with algorithms, and about the nature of influence, given that presumably, algorithms provide informational and not normative influence, that is, people should not feel pressure to get along with the algorithms.

Keywords: Conformity | Algorithm Aversion | Algorithm Appreciation

Word count:

Revisiting Asch: Examining Conformity to Algorithms and Algorithm Aversion

Notes.

- Compare the effects of algorithms to the effect of human crowds. (5 past participants priced this house at x, y, or z.).
- Does it make sense to have a secondary DV of confidence in your score?
- What is the setup: what is the cover story?
- What is the procedure. Do we need to show a couple of correct answers. Should we have a couple of first trials were the algorithms are correct?

In 1956 Solomon Asch published groundbreaking work showing that people will abandon their judgement to conform to a group that is clearly giving the wrong answers to a simple problem (Asch, 1956). Over the following decades, research in social psychology has advanced our understanding of how humans are affected by other humans (Cialdini & Goldstein, 2004). Over the last decades, artificial intelligence and machine learning has advanced such that it now outperfroms humans, even complex tasks. Research in judgment and decision making, has not only compared how human judgment compares to algorithmic judgment (Dawes et al., 1989; Sawyer, 1966), but also how people interact with algorithms. For example, people might distrust an algorithm's prediction of a house price, and prefer to trust a real-estate agent's judgment, a phenomenon dubbed algorithm aversion (Dietvorst et al., 2015).

However, the literature on how people interact with algorithms has relied on testing how people react to a single algorithm, discounting the possibility that when multiple algorithms agree, people might be swayed to conform to them. This possibility raises two conflicting implications: Given that human judgement is often noisy and can be improved by taking algorithms into account (Kahneman et al., 2016; Kahneman et al., 2021), offering an

array of algorithms could improve human judgement by helping people agree with algorithms. On the other hand, algorithms can be biased (e.g., discriminate against subgroups) or lack common sense (e.g., when the input information is uncommon, models may provide nonsensical information). In the context of human-centered artificial intelligence (Riedl, 2019), humans are in the role of supervising algorithms and identyfing potential biases. If people are swayed by algorithms blindly, they might fail to notice algorithmic bias.

We conduct an experiment testing people's propensity to conform to algorithmic recommendations, even when these recommendations are wrong, but come with a group of algorithms that agree with each other. We compare the effect of the number of algorithms, the degree of agreement between algorithms, and the degree to which the algorithms are wrong. Moreover, we compare the relationship between number of sources to confidence when recommendaitons are given by human experts or algorithms.

Conformity

Conformity refers to the adjustment of one's opinions such that they become consistent with the opinions of others or with perceived social norms (*APA Dictionary of Psychology*, n.d.).

The two main drivers of conformity are informational and normative influence (Bernheim, 1994). Informational influence refers to the desire to be correct, and normative influence refers to the desire to be liked and accepted by others (Bernheim, 1994). Conformity is at least partly about status, and, presumably, agreeing or disagreeing with algorithms has no such effect (i.e., only informational influence) (Bernheim, 1994)

Algorithms and human judgement

People and organizations are increasingly interacting with algorithms in their professional (e.g., doctors using diagnostic algorithms) and personal lives (e.g., people deciding on whether to follow traffic directions from google maps; Liel and Zalmanson

(2020)).

Algorithms are generally better than humans, and human judgement amplified by algorithms is still worse than algorithms alone (Kahneman et al., 2021). Algorithms are better: Grove found that 10% better on average (Grove et al., 2000), and even better at recommending jokes (Yeomans et al., 2019).

Algorithms are likely to complement rather than replace human judgemnt (Brynjolfsson & Mitchell, 2017; Shrestha et al., 2019). Thus, it is import to understand how people interact with algorithms. Dietvorst... Logg has found that people prefer algorithms, especially when compared to other people's judgment, rather than teir own; and when they don't think of themselves as experts (Logg et al., 2019). Younger people are more likely to trust algorithms than older people (Kaufmann, 2021).

Liel and Zalmanson (2020) tested people's acceptance of algorithmic suggestions in a simple perceptual task. One limitation of this past research is:

- Uses a task with ambiguous images where subjects must count merged zebras
- Used recommendations of a single algorithm, thus not testing how different algorithms affect judgemnet.
- They measure conformity as providing a wrong answer, but conformity might relate to a quantitative adjusment of opinion that might be milder.
- Their algorithm gave only incorrect responses

The present research

Research on humans interacting with algorithms has focused on two main issues: (1) are there any areas where human judgement is better than algorithms (i.e., clinical vs. actuarial judgment), and (2) are people willing to trust algorithms (i.e., algorithm

aversion and appreciation). The issues of how people react to algorithmic mistakes and how people react to multiple algorithms have been largely ignored.

In this investigation, we conduct an experiment to test how the quantity of algorithms, their level of agreement, and the degree to which the algorithms are wrong affect algorithm appreciation and algorithm aversion. Consistent with literature on human conformity, we predict that more algorithms, that are more in agreement with each other, and are less wrong will result in higher levels of conformity. However, given that presumably people consider algorithms to be more homogeneous than human, we predict that the the effect of the number of algorithms will be weaker than the corresponding effect of the number of humans.

Figure 1

Methods

Participants

We will recruit 4,000 participants from Prolific, and each participant will complete 40 trials, for a total sample size of 40,000 trials.

Materials

Self-report measures

Algorithm appreciation and aversion. We will 4 items out of the algorithm aversion scale (Melick, 2020) use a scale to ascertain baseline levels of trust and mistrust in algorithms, specifically in the financial domain. We hypothesize that people who score high in algorithm aversion will be less likely to conform to algorithms, but not to other people. Items are reproduced in the appendix.

Perceived variance in algorithms. If people react similarly to a single algorithmic recommendation as opposed to multiple, it may be because they believe that

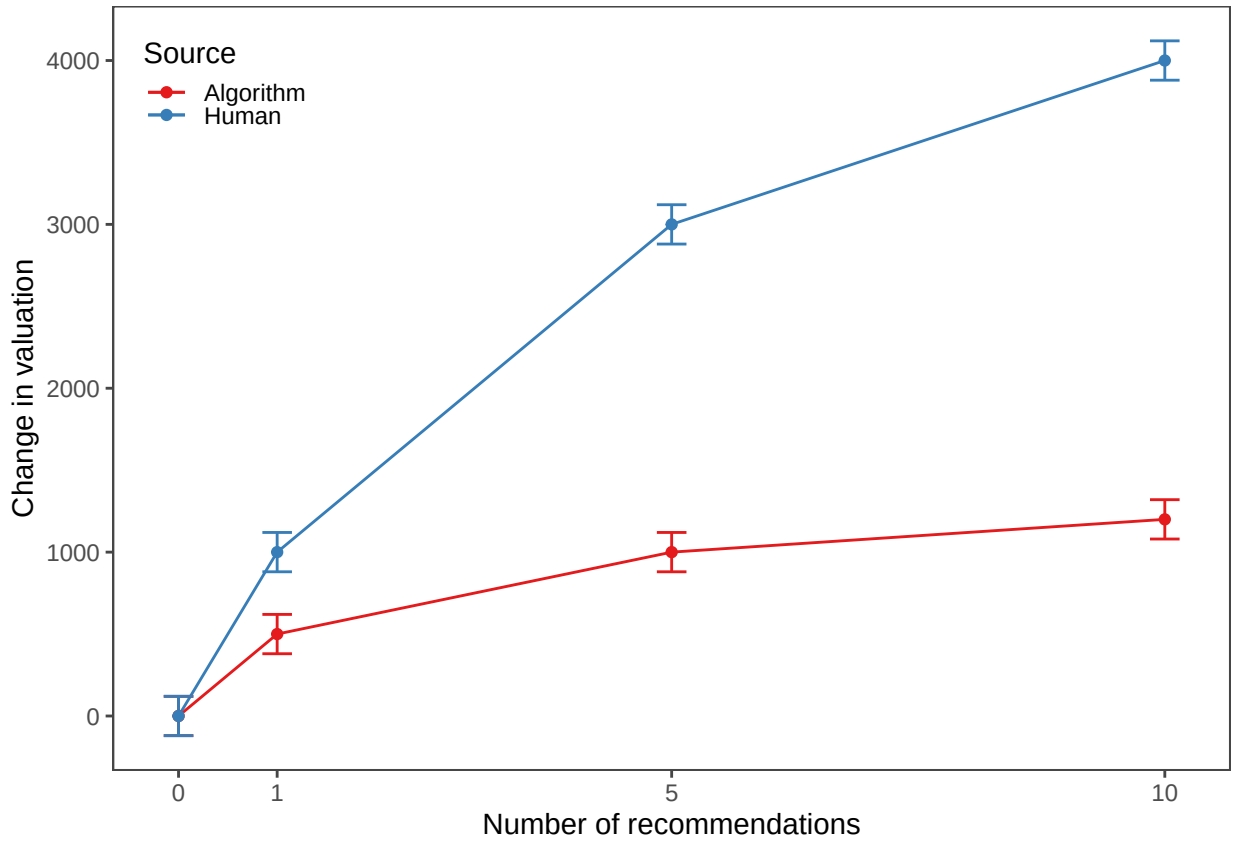

Figure 1

We hypothesize that more recommendations will lead to a larger change in valuation, but the effect of more recommendations will be larger for human recommendations as opposed to algorithmic ones

algorithms tend to be very similar among each other. If people think that algorithmic recommendations are highly correlated (or more correlated to each other than human judgment), then this could explain a potentially flatter relationship between number of recommendations and conformity.

Manipulations

Source of opinions. Across trials, participants will see house price recommendations that are provided by either a number of real estate agents, or a number of different algorithms.

Number of opinions. Across trials, participants will receive recommendations from either 1, 5 or 10 different sources, either all algorithms, or all real-estate agents.

Agreement between opinions. Across trial we will manipulate the level of agreement between the recommendations. In the low-agreement condition, the recommendations will be

Algorithmic bias.

Dependent variable: Change in house price rating

Procedure

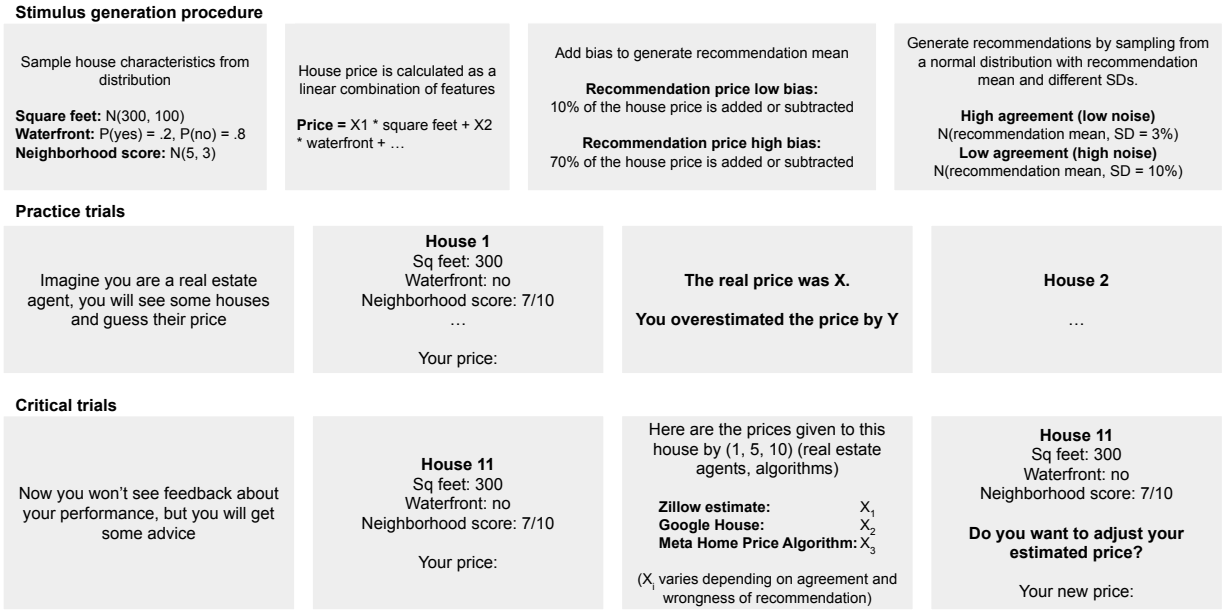

Figure 2

Procedure of the study. Top row shows the stimulus generating procedure, and bottom two rows show the practice and critical trials respectively.

1. Participants read a cover story
2. Participants get a bunch of real house prices so that they can be calibrated See a bunch of examples first to reduce noise in different anchorings.

3. Participants do a number of trials, where they first rate a house price without any algoriths, then see recommendations from either humans or algorithms, and decide whether to change their rating. We will pay people an accuracy bonus to limit concerns of low motivation to be accurate.

Data analysis

We used for all our analyses.

Check for the effect of recognizability

Check for ordering effects

Check for

Results

The result was 8

Discussion

Informational vs. normative influence "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women" (2018)

Limitation: we should do this with real estate people, because people might not consider themselves as experts and therefore might be less prone to algorithm aversion (Logg et al., 2019).

Future research could test the effect of showing human and algorithmic advice concurrently.

References

Amazon scraps secret AI recruiting tool that showed bias against women. (2018). *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

APA Dictionary of Psychology. (n.d.). https://dictionary.apa.org/

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*, 1–70. https://doi.org/10.1037/h0093718

Bernheim, B. D. (1994). A Theory of Conformity. *Journal of Political Economy*, *102*(5), 841–877. https://doi.org/10.1086/261957

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530–1534. https://doi.org/10.1126/science.aap8062

Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, *55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 16681674.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General*, *144*(1), 114–126.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). Noise. How to Overcome the High, Hidden Cost of Inconsistent Decision Making. *Harvard Business Review*, 38–46.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Harper Collins.

Kaufmann, E. (2021). Algorithm appreciation or aversion? Comparing in-service and pre-service teachers' acceptance of computerized expert models. *Computers and Education: Artificial Intelligence*, *2*, 100028. https://doi.org/10.1016/j.caeai.2021.100028

Liel, Y., & Zalmanson, L. (2020). *What if an AI told you that 2 + 2 is 5? Conformity to algorithmic recommendations*.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Melick, S. R. (2020). *Development and validation of a measure of algorithm aversion* [PhD thesis]. https://www.proquest.com/docview/2405300583/abstract/4CA96334EBCA4475PQ/1

Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, *1*(1), 33–36. https://doi.org/10.1002/hbe2.117

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, *66*(3), 178.

Shrestha, Y. R., Ben-Menahem, S. M., & Krogh, G. von. (2019). Organizational Decision-Making Structures in the Age of Artificial Intelligence. *California Management Review*, *61*(4), 66–83. https://doi.org/10.1177/0008125619862257

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414. https://doi.org/10.1002/bdm.2118

Appendix

Algorithmic aversion and appreciation items - Financial domain.

Below are the items for the algorithm aversion scale in the financial domain (Melick, 2020).

- Financial advice that is index-based is more effective than financial advice that is based on the judgment of the advisor. (R)
- It is more appropriate for financial advisors to make recommendations that are indexbased than to make recommendations that are based on their own judgment. (R)
- It is more appropriate for financial lending institutions to make loan decisions based on a mathematical formula designed to predict probability of loan default than based on the judgment of the loan officer. (R)
- Financial lending decisions that are based on the judgment of the loan officer are more effective than lending decisions that are based on the mathematical probability of loan default.