Algorithmic Appreciation Across Task Difficulty

Emergent Research Forum (ERF)

Eric Bogert
University of Georgia
etbogert@uga.edu

Rick Watson
University of Georgia
rwatson@terry.uga.edu

Aaron Schecter
University of Georgia
aschecter@uga.edu

Abstract

We propose a two by two experiment that investigates how humans respond to recommendations based on the difficulty of the task and the source of the recommendation. The two types of information sources which will provide recommendations in our experiment are algorithms and human crowds. We contribute to the burgeoning discourse on algorithmic appreciation by focusing on crowd counting, a task in which effort is a strong factor in predicting accuracy.

Keywords

Algorithmic Appreciation, Algorithmic Aversion, Crowd counting, Task difficulty, Overconfidence

Introduction

There is ample evidence that humans respond to recommendations from both crowds of humans and from algorithms (Castelo et al. 2019; Dietvorst et al. 2015; Logg et al. 2019). Some authors have found that humans prefer other human recommendations (Yeomans et al. 2019), this is *algorithmic aversion*. Algorithmic aversion has been posited as early as the 1950s (Meehl 1954) Others have found the opposite effect (Logg et al. 2019); this is *algorithmic appreciation*. These phenomena are posited to change from one to another based on the perceived objectivity of a task (Castelo et al. 2019). However, we believe that an overlooked concept that could explain these seemingly conflicting results is the difficulty of the task. This paper explores this idea.

Literature Review

Francis Galton introduced the concept of the *wisdom of the crowd* when he found that the average guess of a large group was shockingly accurate (Galton 1907). For crowds to have accurate guesses, three conditions are generally required: 1) a diversity of opinion, 2) opinions are independent, 3) and opinions are decentralized (Surowiecki 2004). The underlying reason these guesses are highly accurate is that each individual's guess combines unique knowledge with some amount of error. In large crowds, the errors cancel out, leaving the average guess highly accurate (Page 2008; Sunstein 2006).

Recent work has investigated the boundaries of the wisdom of the crowd. For example, in tasks with objective answers, such as "What is the length of the border between Switzerland and Italy" when humans are exposed to the estimates of their peers, they tend to become overconfident in their guess, even as their subsequent guesses on the same question become less accurate (Lorenz et al. 2011). This is a well accepted idea – that the average guess should perform worse when the people guessing are exposed to the estimates of their peers (Janis 1972; Lorenz et al. 2011). This occurs because both the diversity of opinion and the independence of opinions, two conditions necessary for wise crowds, decrease once people are shown the guesses of others. Humans have a demonstrated tendency to change their estimates according to the estimates of their peers. What is not clear is whether this propensity is different in large crowds depending on the type of recommendation – human or AI. When humans expend less effort due to some external aid, they are engaging in automation bias (Mosier and Skitka 1996). Automation bias can occur because of aid from fellow humans or aid from AI. There are three sources of automation bias (Parasuraman and Dietrich 2010).

First, the limited rationality of the human brain forces humans to be cognitive misers. Because our cognitive resources are finite, we prefer to not use them, and thus are willing to do that which is less cognitive effort. This has been demonstrated in physiological measures of cognitive effort, showing that heart rates are lower when humans are receiving guidance from AI, compared to when they do not receive guidance (Alexander et al. 2018). Second, some humans show a natural propensity to believe that humans can more successfully assimilate information than they can (Parasuraman and Dietrich 2010). Third, humans engage in social loafing – expending less effort when they are in teams than when they are working alone (Karau and Williams 1993). These factors all contribute to automation bias, and are likely to be affected differently depending on whether humans believe they are paired with a group of AI or a group of humans.

There is a preponderance of evidence that indicates AI are better at well-defined tasks with objective answers than humans are. The best performing financial hedge fund in history runs solely on algorithmic trades (Zuckerman 2019). In a landmark meta-analysis algorithmic judgment was found to outperform human judgment across more than one hundred studies (Grove et al. 2000). The best checkers, chess, and go players are all algorithms (Dockrill 2017; Silver et al. 2017). Despite this, there is ample evidence that humans prefer expert human judgment to AI judgment (Promberger and Baron 2006; Yeomans et al. 2019). This predilection is puzzling, given the superior algorithmic performance. This results in tension. It is possible that humans respond more strongly to the estimates of other humans in a large crowd, given their propensity to prefer expert human judgment. However, it is also possible that the human tendency to prefer human judgment. We expect an important factor in this calculus is the perceived difficulty of the task.

Humans are naturally overconfident (Moore and Healy 2008). This is likely because humans tend to think their own decision processes are more objective than their peers (Liberman et al. 2012). While humans do tend to take the advice of a peer, they usually discount it substantially (Yaniv and Kleinberger 2000). However, in harder tasks, humans have shown stronger tendencies to take advice, although they remain overconfident in their assessment (Gino and Moore 2007).

Research Model & Hypotheses

To test this idea, we propose an experiment in which humans are asked to estimate the number of people in an image. After they submit their guess, they will be exposed to new information. The new information will be the guesses of other people, the guesses of an AI, or the guesses of both other people and AI. In truth, all subjects will receive the same information, only the label, whether it is a guess from an AI or from a human, will change.

Our research model is a 2*2 experiment. The first condition is the advisor source – algorithm or a crowd. The second condition is task difficulty. The easier image will show a quarter of the harder image.

Our research model is shown below in Figure 1.

There are three dependent variables of interest. First, confidence, confidence, measured by a 1-6 on a Likert scale of how confident they are. Confidence will be measured after both the initial guess and the guess after the subject sees the answers of others. Second, belief change, measured using the weight on advice measure (Yaniv and Kleinberger 2000). Third, cognitive effort, measured by the time it takes to submit an answer in the second round.

We believe that average confidence will increase after humans have been exposed to the new information of the guesses from the crowd. This has been demonstrated before in similar tasks (Logg et al. 2019; Lorenz et al. 2011), although our task is substantially different because the answer in our research is plainly available (i.e. the size of the population in an image is actually observable)– whereas in prior research the true answer (i.e. “What is the length of the border between Switzerland and Italy?”) was not directly observable. Fundamentally, we believe that people will be more inclined to believe an algorithm’s recommendation because algorithms are tireless, whereas humans are cognitive misers (Simon 1956). In a task with a directly observable answer, such as counting a crowd, an algorithm's tireless should be a significant strength. Thus, our first hypothesis.

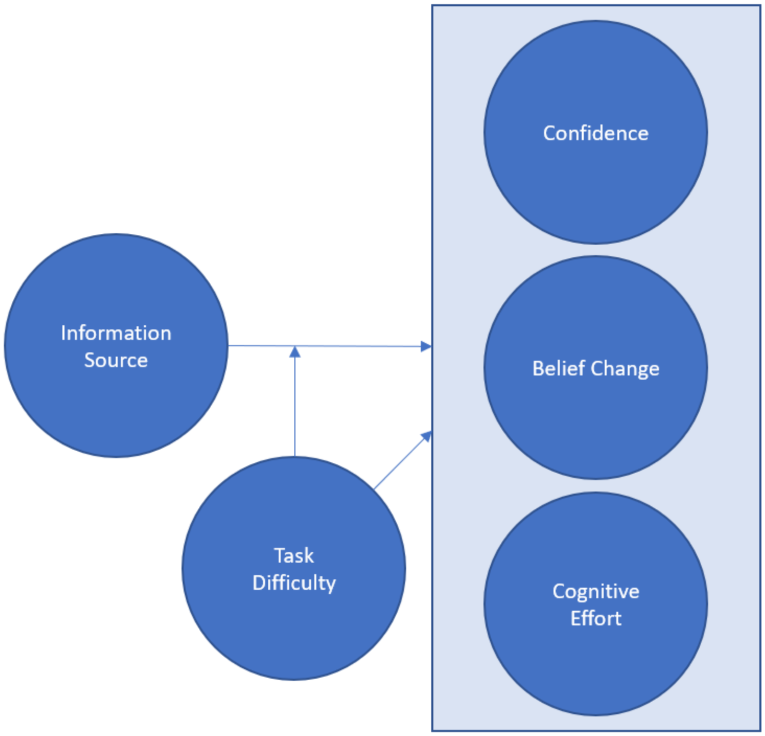

Figure 1: Research Model

Hypothesis 1**:** *Confidence will be higher more when humans are exposed to guesses they believe come from an AI than it will when subjects are exposed to guesses they believe come from other humans.*

We expect that people will be more persuaded by algorithms than by a human crowd. In tasks similar to this one, such as guessing the weight of an individual or estimating how popular a flight is, people have shown algorithmic appreciation (Logg et al. 2019).

Hypothesis 2: *Belief change will be higher when subjects are exposed to guesses they believe come from an AI than it will be when subjects are exposed to guesses they believe come from other humans.*

Algorithmic appreciation should also manifest in cognitive effort. If we observe algorithmic appreciation then we should find that people think less about a problem when they receive an algorithmic recommendation compared to when they receive a recommendation from a human source. Thus, we hypothesize:

Hypothesis 3: *Cognitive effort will be less when subjects are exposed to guesses they believe come from an algorithm than it will be when subjects are exposed to guesses they believe come from other humans.*

Finally, we expect that these effects will all strengthen when subjects are exposed to more difficult questions. This is because the perceived gap in likely effort expended between the algorithm and the average member of the crowd is likely much higher in harder tasks. Thus, we expect the effect of algorithmic appreciation to strengthen in these harder tasks.

Hypothesis 4: *More difficult tasks will strengthen the effect of algorithmic recommendations on cognitive effort, belief change, and confidence.*

Data Collection and Analysis

We will recruit human subjects from Amazon Mechanical Turk (AMT). Amazon Mechanical Turk is an online platform which connects people with Human Intelligence Tasks (HITs). Common HITs on AMT include labeling images, transcription, and filling out surveys. Samples from Amazon Mechanical Turk are transforming academic studies that rely on human participants, because studies on Mechanical Turk are generally cheaper and more convenient (Dance 2015). Samples from Mechanical Turk are usually more diverse than other online samples (Buhrmester et al. 2011; Paolacci and Chandler 2014), and the quality of data produced by individuals from Mechanical Turk can rival engineering graduate students on complex tasks, provided a sufficient tutorial (Staffelbach et al. 2015).

Each subject will be told that they will receive more money the more accurate their estimates are. This is a common technique in studying wisdom of the crowds, because it eliminates any incentive to simply agree with the crowd, which is a strong human tendency (Lorenz et al. 2011).

Although workers on AMT may have systematic differences from other people, we are most interested in how task difficulty changes algorithmic appreciation. Thus, our results should not be contaminated by bias introduced by sampling from AMT as this result is not a point estimation extrapolated to a population (Chandler and Shapiro 2016).

Discussion

This work has significant theoretical and practical implications. It will inform our knowledge of how humans view the utility of a crowd of humans versus a crowd of AI. We build on prior research in several important ways. First and most importantly, we examine the effects of task difficulty. To our knowledge this is a completely unexplored construct in the algorithmic appreciation literature. Second, we use a task that has an observable answer, whereas prior research has questions that rely on an individual's outside knowledge. Third, we are the first paper to look at cognitive effort in the context of algorithmic appreciation.

This paper is also important for practical platform design. For example, when social media uses a fake news flag, it is possible to emphasize whether the flag is driven more by algorithms or by human assessments. If a platform's goal is to maximize the amount of true news that circulated on its platform, it would have be useful to know whether users are more likely to be influenced by being told that crowds of humans versus an algorithm had decided a given article was fake news.

Future work should look at this question in the context of other types of tasks. This study focuses on intellective tasks, tasks with objective, known answers. Other tasks, including creative tasks such as brainstorming or tasks with conflicting viewpoints, such as those with moral components, would expand our knowledge on how AI and human recommendations are task-dependent.

REFERENCES

Buhrmester, M., Kwang, T. N., and Gosling, S. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?," *Perspectives on Psychological Science* (6:1), pp. 3–5.

Castelo, N., Bos, M., and Lehmann, D. 2019. "Task-Dependent Algorithm Aversion," *Journal of Marketing Research* (56:5), pp. 809–825.

Chandler, J., and Shapiro, D. 2016. "Conducting Clinical Research Using Crowdsourced Convenience Samples," *Annual Review of Clinical Psychology* (12), pp. 53–81.

Dance, A. 2015. "News Feature: How Online Studies Are Transforming Psychology Research: The Samples Are Large and Diverse, but Will This Trend Strengthen the Field or Merely Introduce New Sources of Error?," *Proceedings of the National Academy of Sciences* (112:47), pp. 14399–14401.

Dietvorst, B., Simmons, J., and Massey, C. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of Experimental Psychology: General* (144:1).

Dockrill, P. 2017. "In Just 4 Hours, Google's AI Mastered All The Chess Knowledge in History," *Science Alert*.

Galton, F. 1907. "Vox Populi (The Wisdom of the Crowds).," *Nature* (75:1949), pp. 450–451.

Gino, F., and Moore, D. 2007. "Effects of Task Difficulty on Use of Advice," *Journal of Behavioral Decision Making* (20:1), pp. 21–35.

Grove, W., Zald, D., Lebow, B., Snitz, B., and Nelson, C. 2000. "Clinical versus Mechanical Prediction: A Meta-Analysis," *Psychological Assessment* (12:1), pp. 19–30.

Janis, I. 1972. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*, Boston: Houghton, Mifflin.

Karau, S., and Williams, K. 1993. "Social Loafing: A Meta-Analytic Review and Theoretical Integration," *Journal of Personality and Social Psychology* (65:4), pp. 681–706.

Liberman, V., Minson, J., Bryan, C., and Ross, L. 2012. "Naive Realism and Capturing the 'Wisdom of Dyads,'" *Journal of Experimental Social Psychology* (48:2), pp. 507–512.

Logg, J., Minson, J., and Moore, D. 2019. "Algorithmic Appreciation: People Prefer Algorithmic to Human Judgment," *Organizational Behavior and Human Decision Processes* (151), pp. 90–103.

Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. 2011. "How Social Influence Can Undermine the Wisdom of Crowd Effect," *Proceedings of the National Academy of Sciences* (108:22), pp. 9020–9025.

Meehl, P. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, Minneapolis, Minn. : University of Minnesota Press.

Moore, D., and Healy, P. 2008. "The Trouble with Overconfidence," *Psychological Review* (115:2), p. 502.

Mosier, K., and Skitka, L. 1996. "Human Decision Makers and Automated Decision Aids: Made for Each Other?," *Automation and Human Perforamcne: Theory and Applications*.

Page, S. 2008. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*, Princeton: Princeton University Press.

Paolacci, G., and Chandler, J. 2014. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool," *Current Directions in Psychological Science* (23:3), pp. 184–188.

Parasuraman, R., and Dietrich, M. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration," *Human Factors: The Journal of Human Factors and Ergonomics Society* (52:3), pp. 381–410.

Promberger, M., and Baron, J. 2006. "Do Patients Trust Computers?," *Journal of Behavioral Decision Making* (19:5), pp. 455–468.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. 2017. "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm," *ArXiv*.

Simon, H. 1956. "Rational Choice and the Structure of the Environment," *Psychological Review* (63:2), pp. 129–138.

Staffelbach, M., Semploinski, P., Kijewski-Correa, T., Thain, D., Wei, D., Kareem, A., and Madey, G. 2015. "Lessons Learned from Crowdsourcing Complex Engineering Tasks," *PLOS ONE* (10:9).

Sunstein, C. 2006. *Infotopia: How Many Minds Produce Knowledge*, Oxford: Oxford University Press.

Surowiecki. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*, Doubleday;Anchor.

Yaniv, I., and Kleinberger, E. 2000. "Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation," *Organizational Behavior and Human Decision Processes* (83:2), pp. 260–281.

Yeomans, M., Shah, A., Mullainathan, S., and Kleinberg, J. 2019. "Making Sense of Recommendations," *Journal of Behavioral Decision Making* (32:4).