Algorithmic Appreciation in Creative Tasks

Emergent Research Forum (ERF)

Eric Bogert
University of Georgia
etbogert@uga.edu

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

Dr. Aaron Schecter
University of Georgia
aschecter@uga.edu

Abstract

Algorithms have long outperformed humans in tasks with objective answers. In medicine, finance, chess, and other objective fields, AI have been shown to consistently outperform human cognition. However, algorithms currently underperform human cognition in creative tasks, such as writing fiction or brainstorming ideas. We propose a study that investigates how humans rely on algorithmic and human recommendations differently in creative tasks, and whether that effect changes based on task difficulty. We synthesize the current theoretical landscape and propose what the likely effects of algorithmic versus human recommendations are on cognitive effort, belief change, and confidence in output.

Keywords

Artificial Intelligence, McGrath's circumplex model, algorithmic appreciation, algorithmic aversion

Introduction

Algorithms excel at two things: prediction and classification (Goodfellow et al. 2016). This makes them excellent at intellective tasks, those with objective, well-defined answers. Weather forecasting (Silver 2012), maximizing stock market returns (Zuckerman 2019), and playing chess (Dockrill 2017; Silver et al. 2017) are all tasks that algorithms consistently outperform even the best humans in their respective fields. This is because there is an objective function on which to optimize – in the stock market it is returns relative to a benchmark; in chess it's the probability of winning the chess game. However, algorithms are much worse at tasks without a well-defined dependent variable on which to optimize. Although algorithms have been used to write chapters of books (Tewari 2019) and imitate lyrics of popular musicians (Alexander 2017), no one would consider current AI attempts at these creative endeavors to be within the realm of human sophistication. This paper proposes an investigation into how humans respond differently to recommendations for answers to creative tasks, depending on whether they believe the recommendation originates from an algorithm or a human.

Literature Review

When humans receive input from an external source to help them with a task, they are likely to use that input, and as a consequence, expend less cognitive effort (Parasuraman and Riley 1997). This process is called automation bias. Humans are capable of both under-relying and over-relying on external input, and automation bias can be present as a result of fellow human input and non-human input, such as from an AI (Dzindolet et al. 2002). Automation bias occurs because humans are cognitive misers, who are prone to social loafing (Parasuraman and Dietrich 2010). Automation bias towards an AI recommendation relative to a human recommendation is rational in fields in which AI can plausibly outperform humans, such as chess or finance.

Human preferences towards recommendations from other humans has been researched for decades; the original paper documenting the phenomena was written more than 60 years ago (Meehl 1954). Prior research indicates that humans demonstrate preferences toward human recommendations after they have seen an algorithm make a mistake (Dietvorst et al. 2015). This phenomenon is *algorithmic aversion.* In intellective tasks, such as guessing the weight of an individual, people usually display *algorithmic* *appreciation,* a preference toward recommendations from algorithms, when people haven't observed the algorithm make a mistake (Logg et al. 2019). Humans also display more confidence and greater reliance on advice when receiving recommendations from algorithms compared to recommendations from humans in intellective tasks (ibid). Algorithmic appreciation is likely affected by humans' natural propensity to be overconfident.

Overconfidence is extremely common (Kruger and Dunning 1999; Moore and Healy 2008) and makes people rely less on recommendations of peers than they should (Yaniv and Kleinberger 2000). People are overconfident because they believe their analyses are more objective (Liberman et al. 2012). When humans receive advice from a peer, the rational act is to average the advice of the peer with one's own belief, giving the peer's advice a 50% weight. However, because of overconfidence, we usually see a weight of approximately 30% on the advice of the peer (Yaniv and Kleinberger 2000). However, this effect changes based on the difficulty of the task. In harder tasks, people tend to use advice more (Gino and Moore 2007).

We propose to investigate algorithmic appreciation in the context of creative tasks, in which humans currently outperform AI. We differentiate between tasks by use McGrath's Circumplex Model of Group Tasks, see Figure 1. The circumplex model has two types of tasks that we focus on. Intellective tasks, those with objective, correct answers, and creative tasks, those focused on generating ideas.

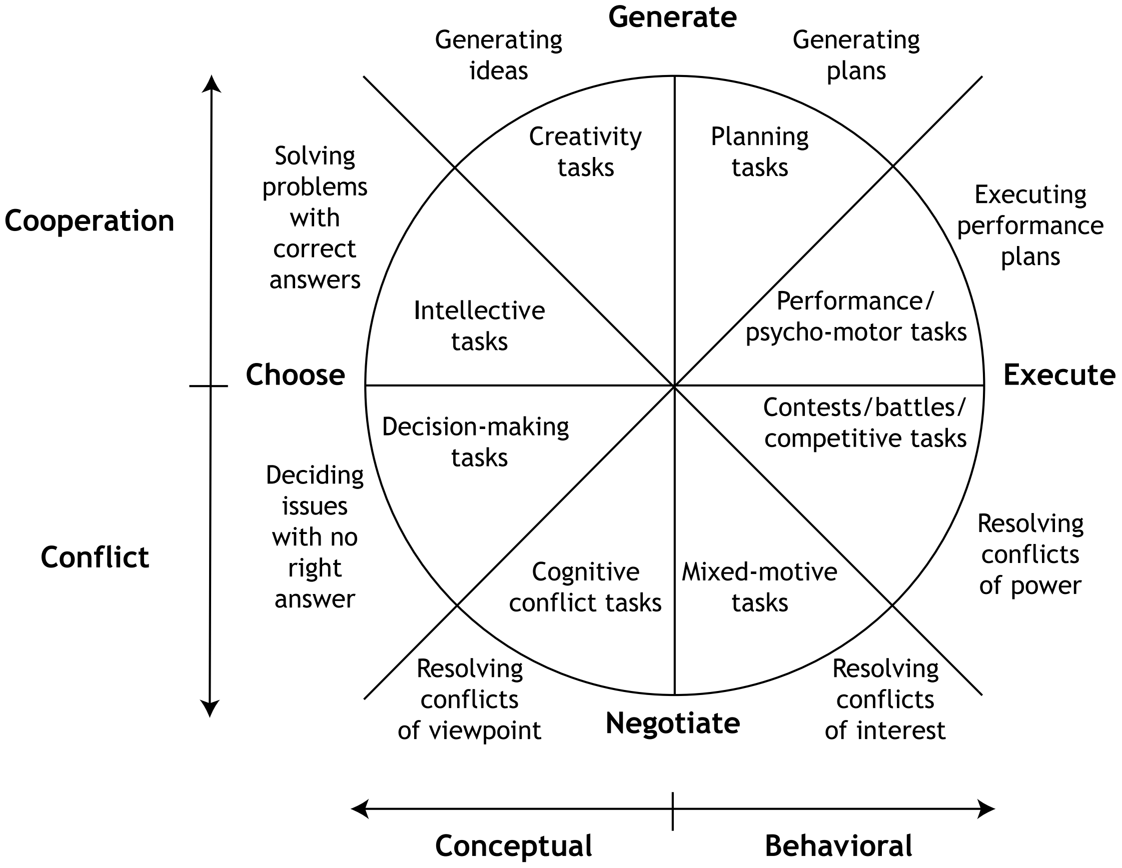
Generate
Generating ideas
Generating plans
Creativity tasks
Planning tasks
Solving problems with correct answers
Executing performance plans
Cooperation
Intellective tasks
Performance/ psycho-motor tasks
Choose
Execute
Decision-making tasks
Contests/battles/ competitive tasks
Deciding issues with no right answer
Resolving conflicts of power
Conflict
Cognitive conflict tasks
Mixed-motive tasks
Resolving conflicts of viewpoint
Negotiate
Resolving conflicts of interest
Conceptual
Behavioral

Figure 1. McGrath's Circumplex Model

Algorithms are excellent at most intellective tasks, assuming large enough datasets on which to train. They are far less good at creative tasks, given the lack of an objective metric on which AI can maximize. There has been some success with algorithms in creative tasks. Algorithms can now generate images of faces that appear human (Kerras et al. 2019) and AI can make online reviews that fool human detectors (Yao et al. 2017). Google and Microsoft have created software that can automatically label certain images, such as landscapes (Kamps 2016; Torbet 2019). However, these steps do not mean algorithms are as good as humans in general purpose creativity tasks.

Propositions

The general propensities to trust recommendations of AI are similar to human propensities to trust recommendations of humans. When an AI has a low observed accuracy, humans are less likely to use its recommendations (Ying et al. 2019). Studies looking at cognitive effort have also observed that humans use more cognitive effort in tasks with AI recommendations when they are told the proportion of humans who also used the AI (Alexander et al. 2018). If humans are aware of the current landscape of AI in creative writing tasks, they may be puzzled at the high quality of these allegedly AI-generated captions. We expect that this will make them think more deeply about the problem. These findings, in conjunction with the knowledge that AI are worse than humans at creative tasks, lead us to our first proposition.

Proposition 1A: *In creative tasks, humans will use more cognitive effort when recommendations are given by an algorithm.*

We are also interested in the interaction between the difficulty of the task and whether the recommendation is from an AI or from a human. This is because most people assume they are smarter than average (Kruger and Dunning 1999). Humans also have a tendency to believe that other humans can outperform AI, because humans can assimilate cues unavailable to the AI, such as social cues (Sundar et al. 2009).

Proposition 1B: *In creative tasks*, t*he effect of the algorithmic recommendation on cognitive effort will be stronger in more difficult tasks.*

If the humans in our study are aware the humorous political cartoon captions are currently beyond the capabilities of even sophisticated AI, they may be somewhat suspicious of our claim that AI generated the caption we show them. We postulate that this suspicion will make our subjects try to differentiate themselves from proposed AI captions more than they try to differentiate themselves from human captions. This informs the degree to which we expect subjects to change their answer. Thus, we propose:

Proposition 2: *In creative tasks, humans will exhibit less belief change when receiving recommendations from an algorithm.*

Finally, we look at the effect on confidence. We expect that in easy tasks subjects will be confident regardless of the type of advisor. This is because the optimal answer in an easy creative task is easy to achieve. Thus, we expect that the type of advisor does not affect confidence in easy tasks. Thus we propose:

Proposition 3A: *In easy creative tasks, humans will be equally confident in their output when they receive recommendations from an algorithm compared to recommendations from a human.*

However, we expect this to change in hard creative tasks, because the optimal answer is far less achievable. In hard creative tasks, such as painting, writing, or teaching, algorithms dramatically underperform humans. AI are notoriously not creative – in pop culture it is common to describe someone who is uncreative as “robotic.” Thus we expect that in hard creative tasks, humans will be more confident when they receive recommendations from a human.

Proposition 3B: *In hard creative tasks, humans will be less confident in their output when they receive recommendations from an algorithm.*

Experimental Design

We propose a 2*2 between-subjects experiment in which we vary task difficulty and whether subjects are told they are receiving a recommendation from an algorithm. The hard tasks will be captioning a political cartoon, the easy tasks will be captioning a landscape. We will tell all subjects that bonuses will be awarded for the top 20% of captions. For easy captions, we will tell subjects their goal should be accuracy, for hard captions, we will tell subjects their goal should be humor. Because we expect all subjects to be able to accurately caption landscapes, we will pay the bonus to anyone who includes a plausible of an image. Each subject will be asked to caption 10 images, five difficult and five easy. After the captioning for all images concludes, we will use a manipulation check, asking them whether their advisor was an algorithm or a human.

The first dependent variable of interest is cognitive effort. We will measure this using the time it takes to generate a caption, a common measure of cognitive effort in MIS (Moravec et al. 2019). The second dependent variable is belief change, the degree to which a subject changes their original caption when they see a recommendation. We will use human raters to judge how different the submitted captions are from the original caption. Third, we will look at how confident each subject is that their caption is funny, using a Likert scale from 1-6.

Creating captions for political cartoons are a quintessential difficult creative task – it is fundamentally generating ideas. We will recruit subjects from Amazon Mechanical Turk (AMT). Subjects will see an image, caption the image, and then be exposed to a recommendation, which will be labeled as generated by either a human or AI. Then, subjects will be asked to generate a new caption for this same political cartoon. We will assess time by taking the total time it takes to write a caption after the individual has arrived at the web page where they can input their caption. We will assess belief change by using human raters to rank the captions by which are most similar to the recommendation. We will assess confidence by using a Likert scale asking how confident the subjects are that their caption is humorous.

Conclusion

We propose an experimental design that assesses whether humans rely differently on the advice from humans compared to the advice of AI. In addition, we assess whether this advice affects cognitive effort, belief change, and confidence. This research is situated in a unique space, as most literature on algorithmic appreciation looks at intellective tasks, where AI should dominate humans, whereas we propose examining the effects on creative tasks, where humans currently outperform AI. Future work should look at how AI and human recommendations affect cognitive effort, quality, confidence, and novelty across other types of tasks in McGrath's Circumplex model. A systematic comparison of how humans respond to AI and human predictions across a variety of tasks could be a fruitful line of inquiry in MIS human-AI research. This study has clear limitations. Our manipulation of whether an experimental subject is exposed to human or AI recommendations could create suspicion that the captions labeled as AI-generated were actually generated by humans, if subjects are aware of the current limitations of AI text generation. We believe that on Amazon Mechanical Turk there are not many people who are aware of the boundaries of AI natural language processing techniques, so we do not expect this to be a fatal flaw.

REFERENCES

Alexander, V. 2017. "Kanye West Rap Verses," *Kaggle*. (https://www.kaggle.com/viccalexander/kanyewestverses, accessed April 2, 2020).

Alexander, V., Blinder, C., and Zak, P. 2018. "Why Trust an Algorithm? Performance, Cognition and Neurophysiology," *Computers in Human Behavior* (89), pp. 279–288.

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*.

Dzindolet, M., Pierce, L., Beck, H., and Dawe, L. 2002. "The Perceived Utility of Human and Automated Aids in a Visual Detection Task," *Human Factors* (44:1), pp. 79–94.

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

Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*, MIT Press.

Kamps, H. J. 2016. "Microsoft Demos Next-Generation Image-Captioning Captionbot," *Techcrunch*. (https://techcrunch.com/2016/03/30/microsoft-caption-bot/, accessed May 2, 2020).

Kerras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. 2019. "Analyzing and Improving the Image Quality of StyleGAN," *ArXiv*.

Kruger, J., and Dunning, D. 1999. "Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments," *Journal of Personality and Social Psychology* (77:6), pp. 1121–1134.

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.

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.

Moravec, P., Minas, R., and Dennis, A. 2019. "Fake News on Social Media: People Believe What They Want to Believe When It Makes No Sense at All," *Management Information Systems Quarterly* (43:4), pp. 1343–1360.

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.

Parasuraman, R., and Riley, V. 1997. "Use, Misuse, Disuse, and Abuse.," *Human Factors: The Journal of Human Factors and Ergonomics Society* (39:2), pp. 230–253.

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*.

Silver, N. 2012. *The Signal and the Noise*.

Sundar, S., Xu, Q., and Oeldorf-Hirsch, A. 2009. "Authority vs. Peer: How Interface Cues Influence Users," in *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009*.

Tewari, U. 2019. "Game of Thrones Episode Script Generation Using LSTM and Recurrent Cells in Tensorflow," *Medium*. (https://medium.com/analytics-vidhya/game-of-thrones-episode-script-generation-using-lstm-and-recurrent-cells-in-tensorflow-c0c40d415a8b, accessed April 2, 2020).

Torbet, G. 2019. "Chrome Will Use AI to Describe Images for Blind and Low-Vision Users," *Engadget*. (https://www.engadget.com/2019/10/10/chrome-image-descriptions-accessibility/, accessed May 2, 2020).

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.

Yao, Y., Viswanath, B., Cryan, J., Zhang, H., and Zhao, B. 2017. "Automated Crowdturfing Attacks and Defenses in Online Review Systems," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1143–1158.

Ying, M., Vaughan, J. W., and Wallach, H. 2019. "Understanding the Effect of Accuracy on Trust in Machine Learning Models," *Conference on Human Factors in Computing Systems*, pp. 1–12.

Zuckerman, G. 2019. *The Man Who Solved the Market*, New York, NY: Portfolio Penguin.