CellPress

Trends in
Cognitive Sciences

Science & Society

Preference for human, not algorithm aversion

Carey K. Morewedge [1,*,@]

People sometimes exhibit a costly preference for humans relative to algorithms, which is often defined as a domain-general algorithm aversion. I propose it is instead driven by biased evaluations of the self and other humans, which occurs more narrowly in domains where identity is threatened and when evaluative criteria are ambiguous.

Algorithms outperform humans in many domains and are forecasted to replace as much as 47% of US employment over the next two decades. Algorithms already exhibit comparable or superior performance to humans in tasks ranging from finance to fraud detection to oncology to poetry to paralegal work [1]. Still, people exhibit a costly preference to rely on themselves or other humans relative to algorithms for a variety of judgments, decisions, and services, a preference often defined as algorithm aversion [2]. This definition implies the preference is general and driven by a prejudiced perception of algorithms. Relative to human judges and experts, it suggests people judge the performance of algorithms more critically, perceive their decision processes to be more opaque and rigid, are more sensitive to the mistakes of algorithms, and perceive algorithms as less capable of learning and assessing qualitative and moral attributes [3–5]. This framing counsels humanizing or reducing prejudice toward algorithms as paths toward increasing their adoption (e.g., via anthropomorphic interventions) [2,5] and guaranteeing human oversight for cases where prejudiced views of algorithms are intractable [5,6].

I propose that defining the preference as an aversion to algorithms misinterprets the directionality of the effect, its breadth, the underlying cause of the preference, and impedes the design of interventions to facilitate the adoption of algorithms. By most objective measures, people perceive algorithms fairly. People claim to have a limited understanding of algorithmic decision processes, and they are right [4]. Laypeople recognize that algorithms perform better than themselves, perceive algorithms to be less biased than humans, and readily outsource decisions to algorithms in many domains (e.g., information, finances, shopping, entertainment, dating) [7,8]. When controlling for confounds like differences in the performance, normativity, novelty, or convenience of algorithms and individual differences within humans in their adoption of new technologies, there is little prejudice toward algorithms to abate.

I suggest that the preference for humans relative to algorithms is more parsimoniously explained by biased evaluations of the self and, by extension, other humans. Like other social judgments, I propose people evaluate algorithms by comparison to the self and the social groups to which they belong. Research on biased self-evaluation finds people exhibit self-enhancing and self-protective biases when unfavorable comparisons in domains central to their identity evoke negative emotions and threaten their self-esteem [9]. This theory predicts that comparisons between humans and algorithms should be unbiased in domains that are peripheral to the identity of the (human) judge. In domains that are central to their identity, however, threatening comparisons may evoke self-enhancing and self-protective biases. As evidence, people overestimate their own performance and abilities, underestimate the performance and abilities of algorithms, and emphasize dimensions of comparison that are favorable to themselves and other humans or that are unfavorable to algorithms (e.g., feelings of trust rather than objective performance) [2,5,10]. People are more likely to overestimate their understanding of human than algorithmic decision processes and overlook human biases unless explicitly reminded [4,11]. Indeed, the term 'algorithmic bias' is used to describe biases in predictions made by algorithms, despite their human origin. Biases in algorithms usually reflect the biased human decisions on which they are trained, the data and statistical assumptions humans choose to include or exclude in their training, or human decisions to inappropriately use algorithmic predictions for outcomes and in contexts to which they do not apply [12].

Reframing this phenomenon as a preference for humans driven by biased evaluations of self and other humans reconciles past findings and yields novel predictions for future research. This theoretical framework suggests the preference should be driven by two factors modulating bias in self-evaluations [9]: (i) the identity-relevance of the judgment, decision, skill, or task and (ii) the ambiguity of its evaluative criteria (*x*-axis and *y*-axis, Figure 1). People should exhibit a stronger preference for humans than algorithms in domains that are relevant to their identity and when evaluative criteria are ambiguous. As identity-relevance and the ambiguity of evaluative criteria decrease, so should the strength of the preference for humans. In domains irrelevant to their identity with clear evaluative criteria, people should be indifferent between equivalent humans and algorithms. They may even exhibit algorithm appreciation [8] – prefer algorithms that are superior on orthogonal attributes (e.g., cost, accuracy, convenience, normativity) or reduce their identification with undesirable activities (e.g., stigmatized, unethical).

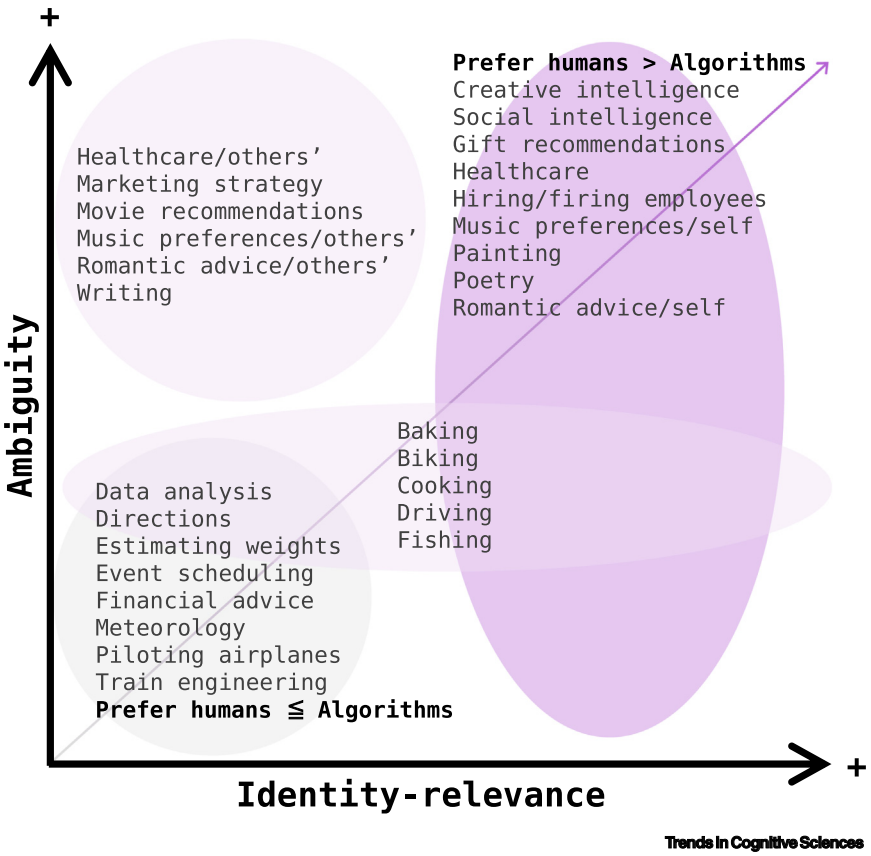

Figure 1. Preferences for humans relative to algorithms by identity-relevance and ambiguity of evaluative criteria. If the preference for humans relative to algorithms is driven by biased self-evaluations, it should increase with the (i) identity-relevance of the judgment, decision, skill, or task (*x*-axis) and (ii) ambiguity of its evaluative criteria (e.g., data and process it requires; *y*-axis). Quadrants are populated with examples from references. Shaded area suggests *x* and *y* values of clusters are variable; the intensity of color saturation suggests general preference strength in clusters.

Evidence supporting the identity-relevance prediction includes people's preference to rely on their own judgments, decisions, and actions rather than algorithms and automation in domains where they are experts [8] or are otherwise important to their identity. People who perceive driving to be an important facet of their identity are less willing to adopt autonomous vehicles than people who do not identify with driving. Adoption of automation is also modulated by identification with other activities like baking, biking, cooking, and fishing [13]. Furthermore, people prefer to rely on human rather than algorithmic judgments and decisions for capacities they view as uniquely human, such as moral judgment [14], and in domains in which they perceive themselves to be unique, such as their health [11]. In identity-relevant domains, people also exhibit a stronger preference for relying on judgments and decisions made by other humans than algorithms for themselves (cf. but not for other people). Preferences for human relative to algorithmic healthcare providers increase with the extent to which patients perceive their health to be unique [5]. However, whereas most patients prefer to receive their own healthcare from a human provider [4,10], they exhibit no preference for whether humans or algorithms provide healthcare to other (deindividuated) patients [5].

Evidence supporting the evaluative criteria prediction includes people's stronger preference for humans than algorithms in domains where evaluative criteria are perceived to be more subjective than objective. Facebook users click more on ads for dating advice provided by humans than algorithms (a more subjective domain) but click similarly on ads for financial advice provided by humans or algorithms (a more objective domain) [2]. Describing the same decision as more objective than subjective (e.g., benefiting from quantitative analysis vs. intuition) or explaining its objective criteria also reduces the preference for humans relative to algorithms [2,4].

This theoretical framework proposes that paths to facilitate the adoption of algorithms include reducing the identity-threat they evoke and increasing the perceived objectivity of evaluative criteria. Interventions that make tasks less identity-relevant or buffer identity-threats (e.g., self-affirmations [9]) should reduce preferences for humans. More broadly, the framework predicts in which domains people will prefer humans to algorithms. At a population level, people identify with some preferences and abilities more than others (e.g., 'symbolic' than utilitarian preferences) [13]. At the individual level, people identify more with preferences and abilities for which they possess expertise or that are distinctive associations with desired social groups [9,13]. At least half of the occupations that social scientists believe are most difficult to automate, for instance, rely on abilities central to academic expertise in the social sciences (e.g., creative and social intelligence) [1]. Algorithms have changed and will continue to change our lives for the better (Box 1), except when our fragile identities lead us to reject their benefits in the domains that we care about most.

Declaration of interests

No interests are declared.

Box 1. Should people be threatened by algorithms?

Concern over technological unemployment has been fomented with each technological revolution, often by those benefiting from the status quo. Labor-economizing technologies do have destructive effects, increasing inequity when labor reallocation is concentrated among the least skilled. Their corresponding productivity increases, however, reduce prices, increase income, and expand employment elsewhere. Like other technologies, algorithms will reduce human employment in many industries. Forecasts suggest jobs requiring complex perception and manipulation, creative intelligence, and social intelligence will be replaced later due to challenges in engineering and articulating tasks they entail [1]. Algorithms not only replace jobs. Algorithms facilitate nonroutine human judgments and decisions. In journalism, algorithms perform many routine tasks (e.g., moderating comments, headline A/B testing, writing basic earnings reports). Rather than costing jobs, this has facilitated investigative journalism. Algorithmic labor gives journalists time to analyze information, conduct interviews, and discern what to investigate. Algorithms also support complex investigations. Algorithms were crucial for scanning, compiling, and analyzing the 11.5 million leaked documents comprising the Panama Papers. Without algorithms, it would have been nearly impossible to process the 2.6 terabytes of data on which the investigative reports were based [15].

[1]Department of Marketing, Questrom School of Business, Boston University, Boston, MA 02215, USA

*Correspondence:
morewedg@bu.edu (C.K. Morewedge).
@Twitter: @morewedge

https://doi.org/10.1016/j.tics.2022.07.007

References

1. Frey, C. and Osborne, M. (2017) The future of employment: how susceptible are jobs to computerization? *Technol. Forecast. Soc. Change* 114, 254–280
2. Castello, N. *et al.* (2019) Task-dependent algorithm aversion. *J. Mark. Res.* 56, 809–825
3. Dietvorst, B.J. *et al.* (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* 144, 114–126
4. Cadario, R. *et al.* (2021) Understanding, explaining, and utilizing medical artificial intelligence. *Nat. Hum. Behav.* 5, 1636–1642
5. Longoni, C. *et al.* (2019) Resistance to medical artificial intelligence. *J. Consum. Res.* 46, 629–650
6. Dietvorst, B.J. *et al.* (2018) Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. *Manag. Sci.* 65, 1155–1170
7. Demirdag, I. and Shu, S. (2021) Bias neglect: when human bias, but not algorithmic bias, is disregarded. In *NA - Advances in Consumer Research* (Bradford, T.W. *et al.*, eds), pp. 160–161, Association for Consumer Research
8. Logg, J.M. *et al.* (2019) Algorithm appreciation: people prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* 151, 90–103
9. Sedikides, C. *et al.* (2021) On the utility of the self in social perception: an egocentric tactician model. In *Advances in Experimental Social Psychology*, pp. 247–298, Academic Press
10. Promberger, M. and Baron, J. (2006) Do patients trust computers? *J. Behav. Decis. Mak.* 19, 455–468
11. Bigman, Y.E. *et al.* (2021) Threat of racial and economic inequality increases preference for algorithm decision-making. *Comput. Hum. Behav.* 122, 106859
12. Kleinberg, J. *et al.* (2018) Discrimination in the age of algorithms. *J. Leg. Anal.* 10, 113–174
13. Leung, E. *et al.* (2018) Man versus machine: resisting automation in identity-based consumer behavior. *J. Mark. Res.* 55, 818–831
14. Bigman, Y.E. and Gray, K. (2018) People are averse to machines making moral decisions. *Cognition* 181, 21–34
15. Diakopoulos, N. (2019) *Automating the News: How Algorithms are Rewriting the Media*, Harvard University Press