Judgment and Decision Making, Vol. 15, No. 3, May 2020, pp. 449–451

Algorithm aversion is too often presented as though it were non-compensatory: A reply to Longoni et al. (2020)

Mark V. Pezzo[*] Jason W. Beckstead[†]

Abstract

We clarify two points made in our commentary (Pezzo & Beckstead, 2020, this issue) on a recent paper by Longoni, Bonezzi, and Morewedge (2019). In both Experiments 1 and 4 from their paper, it is not possible to determine whether accuracy can compensate for algorithm aversion. Experiments 3A-C, however, do show a strong effect of accuracy such that AI that is superior to a human provider is embraced by patients. Many papers, including Longoni et al. tend to minimize the role of this compensatory process, apparently because it seems obvious to the authors (Longoni, Bonezzi, Morewedge, 2020, this issue). Such minimization, however, can lead to (mis)citations in which research that clearly demonstrates a compensatory role of AI accuracy is cited as non-compensatory.

Keywords: automation, artificial intelligence, healthcare, uniqueness, medical decision making, trust, decision aids

1 Introduction

We thank the editor for this opportunity to clarify the position we (Pezzo & Beckstead, 2020, this issue) took in our commentary on Longoni, Bonezzi, and Morewedge (2019). To restate, we believe that Longoni et al performed an excellent series of experiments highlighting an important construct – uniqueness neglect – that holds great promise in explaining resistance to AI. We welcome the clarification by Logoni et al. (2020, this issue) that they do not subscribe to a non-compensatory decision process. We appreciate that Longoni et al. (2019) were not particularly interested in cases in which AI was superior to the human, and so their paper did not highlight the compensatory aspect of the model. We wrote the commentary, however, because we believe that many readers would be interested in this aspect, and that Logoni et al. had very interesting data that addressed it.

It is important to note that algorithm aversion is typically introduced as though it were non-compensatory – at least concerning accuracy. Most authors introduce the topic by providing numerous examples of aversion to artificial intelligence even when its accuracy is superior to that of a human judge (e.g., Dietvorst, Simmons & Massey, 2018). Longoni et al begin their paper with two such examples (Donnelly, 2017; Lohr, 2016) and we respectfully maintain that some key statements in their paper could be easily misinterpreted as saying they found resistance to AI even when it was more accurate, despite the inclusion of other, more subtle statements to the contrary. As a result, to our knowledge none of the 25+ articles citing Longoni et al. to date have mentioned the important caveat that resistance only occurs when AI and Human are equal in accuracy. A few have gone so far as to explicitly – and incorrectly – cite Longoni et al. as evidence that algorithm aversion occurs even when the AI is more accurate (Carmon, Schrift, Wertenbroch & Yang, 2019; Páez, 2020). As one reviewer of our original commentary noted, such mis-readings are not uncommon. For example, Dietvorst et al. (2015) showed that preference for a human occurred only after seeing the algorithm err. Those in a control condition, however, actually preferred the algorithm over their own or others' judgments. Logg, Minson and Moore (2019) noted that this paper, nevertheless, has been cited multiple times as a form of non-compensatory algorithm aversion. Thus, a commentary seems the perfect opportunity to clarify and avoid such misunderstandings. With this in mind, we offer two additional clarifications.

First, in their reply to our commentary Longoni et al. (2020) state that it was "obvious" to them (p. 3) that informing participants of AI's superior accuracy would compensate for algorithm aversion, however they acknowledge that it may not have been so to other readers. We agree that it is not obvious to most readers, both for reasons we stated earlier and because the very existence of uniqueness neglect reported by Longoni et al. (2019) implies a *distrust* of reported accuracy levels. That is to say, even when AI has been presented as (historically) more accurate than human, it is easy to imagine that some people might still prefer the human because they imagine themselves as unique and thus outside of the parameters of the algorithm used by the computer. The superior accuracy of AI may not be enough to satisfy individuals

[*] Department of Psychology, University of South Florida St. Petersburg. Email: pezzo@usf.edu. ORCID 0000-0002-4442-6244.

[†] College of Public Health, University of South Florida Tampa. ORCID 0000-0002-7599-2556.

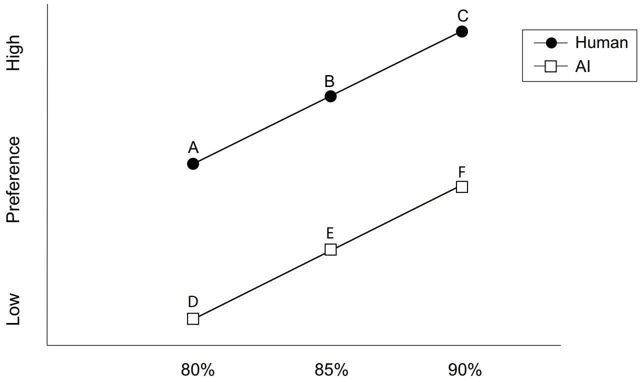

FIGURE 1A. Hypothetical data showing a main effect of provider and accuracy on preference. Here, the AI is never preferred, regardless of relative accuracy.

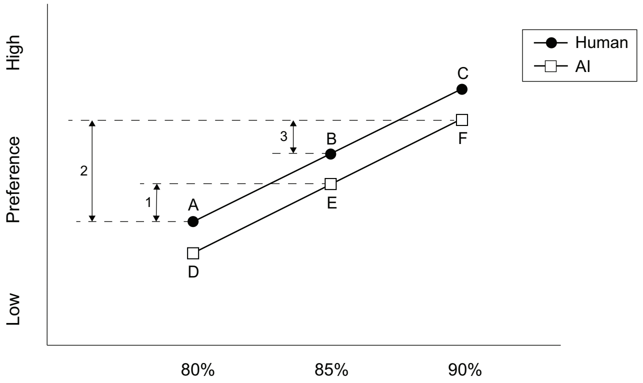

FIGURE 1B. Hypothetical data showing a main effect of provider and accuracy, in which AI is preferred when it has superior accuracy to that of human provider. Comparison 1: $E_{AI} > A_{Human}$; Comparison 2: $F_{AI} > A_{Human}$; Comparison 3: $F_{AI} > B_{Human}$.

scoring high on fears of uniqueness neglect.

Second, we should clarify why we characterized Experiments 1 and 4 as "not allow[ing] for a direct comparison between human and computer" (p. XX). In Experiment 1 any given participant received information only about the human provider, or about the AI provider, but never both. Thus, although the study design permits the analysts to compare provider types, it does not offer participants the opportunity to do so. Further, because the accuracy levels provided for human and AI were always equal, Experiment 1 does not address, nor does it contradict our point.

Regarding Experiment 4, perhaps we should have said that it did not allow for a complete comparison between AI and human providers, as the experiment did not utilize a full factorial design. While the fractional factorial design employed did permit unbiased estimates of (dis)utilities at the aggregate level, the design did not require each participant to respond to all 2 x 3 x 3 = 18 condition combinations, but only to a subset of 7, so direct analytical comparisons of cell means are not possible. Such comparisons are critical to determine if accuracy can compensate for algorithm aversion. If such comparisons had been performed, we can imagine two possible outcomes, one that is compensatory, and one that is not, as shown in Figures 1A and 1B.

Figure 1A depicts *hypothetical* data for the 2 x 3 (provider type by accuracy level) factorial design at the center of our discussion. Similar to Longoni et al. (2019) there is a main effect of both provider type and accuracy level. In this example, the human provider is always preferred, regardless of AI accuracy. All values for AI (points D, E, and F) fall below the lowest value for the human provider (point A). Thus, Figure 1A represents an apparent *non-compensatory* result.

Figure 1B depicts the same hypothetical data with a subtle but important difference; now, points E and F do not fall below point A. Again, main effects of provider type and accuracy level exist, but here the main effect of provider is smaller. As a result, when AI has superior accuracy to the human it is actually preferred. This may be shown by three contrasts applied to pairs of means. Contrast 1 (points A vs. E) compares preference for Human and AI when the accuracy of the AI (85%) is somewhat better than that of the human provider (80%). Contrast 2 (points A vs. F) compares preference for human and AI when the accuracy of the AI (90%) is considerably better than that of the human provider (80%). Contrast 3 (points B vs. F) is similar to Contrast 1 in that it again compares preferences when the accuracy of the AI (90%) is somewhat better than that of the human provider (85%). Figure 1B thus represents a clear *compensatory* result. Algorithm aversion still exists, but may be offset by increasing the relative accuracy of AI.

Algorithms, and AI in particular have been extremely promising as an effective way to provide safe, reliable, and cost-effective medical care. As noted elsewhere (Pezzo, Nash, Vieux & Foster-Grammer, 2020) not all research demonstrating algorithm aversion has provided the sort of detailed accuracy information that Longoni et al. (2019) have. When such information is not provided, Arkes (2008) suggests that people likely assume that computers are not as accurate as humans. The good news is that computers are usually better (Grove, Zald, Lebow, Snitz & Nelson, 2000), and that people seem willing to embrace AI when they are told (and believe) this. Of course, whether people believe the accuracy data they receive may be determined by the extent to which people view themselves as unique as Longoni et al. have shown. This is an exciting direction for future research.

2 References

Arkes, H. R. (2008). Being an advocate for linear models of judgment is not an easy life. In R. M. Dawes & J. I. Krueger (Eds.), *Rationality and social responsibility: Essays in honor of Robyn Mason Dawes* (pp. 47–70). CRC Press.

Carmon, Z., Schrift, R., Wertenbroch, K., & Yang, H. (2019). Designing AI systems that customers won't hate. *MITSloan Management Review*, (Reprint #61315), https://mitsmr.com/2qY8i35.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

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*(1), 19–30.

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.

Longoni, C., Bonezzi, A., & Morewedge, C. (2020). Resistance to medical artificial intelligence is an attribute in a compensatory decision process: Response to Pezzo and Beckstead (2020). *Judgment and Decision Making, 15*(3), 446–448.

Longoni, C., Bonezzi, A., & Morewedge, C. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Psychology*, *46*(4), 629–650.

Páez, A. (2020). Robot mindreading and the problem of trust. Retrieved from http://arxiv.org/abs/2003.01238.

Pezzo, M. V., & Beckstead, J. W. (2020). Patients prefer AI to humans, so long as the AI is better than the humans: A commentary on Longoni, Bonezzi, and Morewedge (2019). *Judgment and Decision Making, 15*(3), 449–451.

Pezzo, M. V., Nash, E. B., Vieux, P. T., & Foster-Grammer, H. W. (2020). The effect of having, but not consulting a computerized diagnostic aid on physician perceptions. *Manuscript Under Review*.