AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2022 American Psychological Association
ISSN: 0096-3445

2022, Vol. 151, No. 9, 2250–2258
https://doi.org/10.1037/xge0001181

BRIEF REPORT

The Human Black-Box: The Illusion of Understanding Human Better Than Algorithmic Decision-Making

Andrea Bonezzi[1], Massimiliano Ostinelli[2], and Johann Melzner[1]
[1] Stern School of Business, New York University
[2] College of Business Administration, Winthrop University

As algorithms increasingly replace human decision-makers, concerns have been voiced about the black-box nature of algorithmic decision-making. These concerns raise an apparent paradox. In many cases, human decision-makers are just as much of a black-box as the algorithms that are meant to replace them. Yet, the inscrutability of human decision-making seems to raise fewer concerns. We suggest that one of the reasons for this paradox is that people foster an illusion of understanding human better than algorithmic decision-making, when in fact, both are black-boxes. We further propose that this occurs, at least in part, because people project their own intuitive understanding of a decision-making process more onto other humans than onto algorithms, and as a result, believe that they understand human better than algorithmic decision-making, when in fact, this is merely an illusion.

Keywords: understanding, projection, illusion of explanatory depth, algorithms, algorithm aversion

Supplemental materials: https://doi.org/10.1037/xge0001181.supp

In 2013, Eric Loomis pled guilty to operating a vehicle without the owner's consent and was sentenced to 6 years of detention. The length of the sentence was based, in part, on the results of a risk assessment conducted by an algorithm that deemed him at high risk of recidivism (Smith, 2016). Loomis appealed the sentence on the ground that the proprietary nature of the algorithm precluded understanding how it judged him at high risk of recidivism. In essence, Loomis took exception to the black-box nature of the algorithm.

Concerns about the black-box nature of algorithms have been raised in several other domains, such as hiring and health care (Campolo et al., 2017). Experts have been discussing the need to make the inner-workings of algorithms transparent (Watson & Nations, 2019), and policymakers have implemented laws that establish people's right to understand how algorithms work (Goodman & Flaxman, 2017). For example, some state legislation requires companies that use algorithms in hiring to disclose how an algorithm evaluates candidates (Artificial Intelligence Video Interview Act, 2020).

The emphasis on making algorithmic decision-making transparent, although well-motivated, raises a paradox. Every day judges evaluate defendants without explaining how they arrive at their judgments (Vigorita, 2003), recruiters make hiring decisions without explaining how they evaluate candidates (Klehe et al., 2008), and physicians make diagnoses without explaining how to patients (Mangano et al., 2015). As these examples illustrate, human decision-makers are often just as much of a black-box as the algorithms that are meant to replace them. Yet, the inscrutability of human decision-making seems to raise fewer concerns. Neither judges nor recruiters or physicians are under obligation to explain how they make decisions (Cohen, 2015; Estlund, 2010; Murray, 2012) and the results of three experiments suggest that people demand less transparency from human than from algorithmic decision-makers (see online supplemental materials A).

We propose that one of the reasons for this paradox is that people foster the illusion to understand human better than algorithmic decision-making, when in fact, both are black-boxes. Prior research shows that people often overestimate how much they understand how things work, a phenomenon referred to as the illusion of explanatory depth (IOED; Rozenblit & Keil, 2002). The IOED has been documented for mechanical devices, natural phenomena, and public policies (Fernbach et al., 2013; Keil et al., 2004; Mills & Keil, 2004). We propose that such IOED also applies to decision-

This article was published Online First February 10, 2022.
Andrea Bonezzi https://orcid.org/0000-0002-5624-8406
None of the data appearing in this article have been previously disseminated in any form (e.g., conferences, meetings, posted on a listserv or website, etc.). An earlier version of the idea advanced in this article was presented at the Association for Consumer Research Annual Conference in 2020. Materials, data and analysis code for all experiments are available on OSF (https://osf.io/9tqez; Bonezzi et al., 2021).

Correspondence concerning this article should be addressed to Andrea Bonezzi, Stern School of Business, New York University, 40 West 4th Street, New York, NY 10012, United States. Email: abonezzi@stern.nyu.edu

making in that people foster an illusion of understanding how other agents make judgments and decisions. We further propose that such an illusion is stronger for human than for algorithmic agents.

We argue that one of the reasons why this occurs is that people project their own intuitive understanding of a decision-making process more onto other humans than onto algorithms. Our prediction builds on prior research suggesting that people often attempt to "read" the mind of others by projecting their own cognitions onto those others (Allport, 1924; Ames, 2005; Krueger, 1998). Projection has been conceptualized as a process of egocentric anchoring whereby judgments about others are anchored on one's own introspections (Krueger, 2000). People intuit what others think, feel, or do in a certain situation by projecting onto others their own thoughts, feelings, and preferences in that situation (e.g., Ross et al., 1977; Van Boven & Loewenstein, 2003; for a review see Krueger, 2007).

The extent to which people project their own mental states onto others is moderated by the perceived similarity with the target: the more similar to the self others are perceived to be, the more people project onto them; as perceived similarity decreases, so does the extent to which people project onto others (Ames, 2004a, 2004b). Ample evidence across a variety of targets, manipulations of similarity, and dimensions of projection supports this notion (e.g., Ames et al., 2012; Davis, 2017; O'Brien & Ellsworth, 2012; Tamir & Mitchell, 2013). For example, much evidence shows that projection is facilitated by shared group membership (Clement & Krueger, 2002; Krueger & Zeiger, 1993; for a meta-analytic review see Robbins & Krueger, 2005) because shared group membership increases perceived similarity with the target (Woo & Mitchell, 2020). And neuroimaging research shows that neural regions associated with self-referential thought are more activated when people mentalize about similar than dissimilar others, suggesting that people recruit information about the self to intuit the mental states of similar more than dissimilar targets (Mitchell et al., 2005, 2006).

Drawing on this literature, we propose that because people are more similar to other humans than to algorithms (Epley et al., 2007; Gray et al., 2007; Haslam, 2006), they are more likely to rely on their own understanding of a decision-making process to intuit how other humans, versus algorithms, make decisions. The privileged—yet often misguided—view that projection provides into other humans' minds can foster the illusion of understanding human better than algorithmic decision processes, when in fact, both are black-boxes.

Six experiments test our hypotheses. Experiments 1A–C test whether people foster a stronger illusion of understanding human than algorithmic decision-making across three domains. Experiments 2, 3, and 4 (in online supplemental materials E) test whether projection accounts for this phenomenon in each domain. Experiment 4 also tests how illusory understanding affects trust in human versus algorithmic decisions. New York University and Winthrop University Institutional Review Board (IRB) approved the experimental protocols. In all experiments, the sample size was predetermined, and a sensitivity power analysis (Faul et al., 2009) indicated that small-to-medium size effects could be detected with a power of .80. We report all conditions, manipulations, measures, and data exclusions. Questions to screen for bots and avoid differential dropout were included at the beginning of each experiment (see online supplemental materials B).

Experiments 1A–C

Experiments 1A–C test whether people foster a stronger illusion of understanding human than algorithmic decision-making. We examined three domains of key societal relevance, where algorithms are increasingly replacing human decision-making—criminal justice (1A), recruiting (1B), and health care (1C)—and modeled our stimuli after existing applications. We leveraged the classic IOED paradigm whereby illusory understanding is revealed by asking people to explain in detail how something works (Rozenblit & Keil, 2002). We predicted that respondents would think that they understand human better than algorithmic decision-making unless prompted to explain the decision-making process in detail.

Method

Respondents recruited from Amazon Mechanical Turk (MTurk) were randomly assigned to a 2 (decision-maker: human, algorithm) × 2 (explanation: yes, no) between-subjects design. Respondents were asked to consider how a judge/algorithm evaluates a defendant's risk of recidivism (1A), how a recruiter/algorithm examines video interviews to evaluate applicants (1B), and how a radiologist/algorithm examines magnetic resonance imaging (MRI) images to diagnose a disease (1C). In the explanation condition, participants were first prompted to explain the decision-making process of the human/algorithm, then rated their understanding (1 = *do not understand* at all to 7 = *completely understand*). In the no-explanation condition, respondents rated their understanding, then were prompted to explain the decision-making process to avoid differential dropout (see Appendix for details).

Results

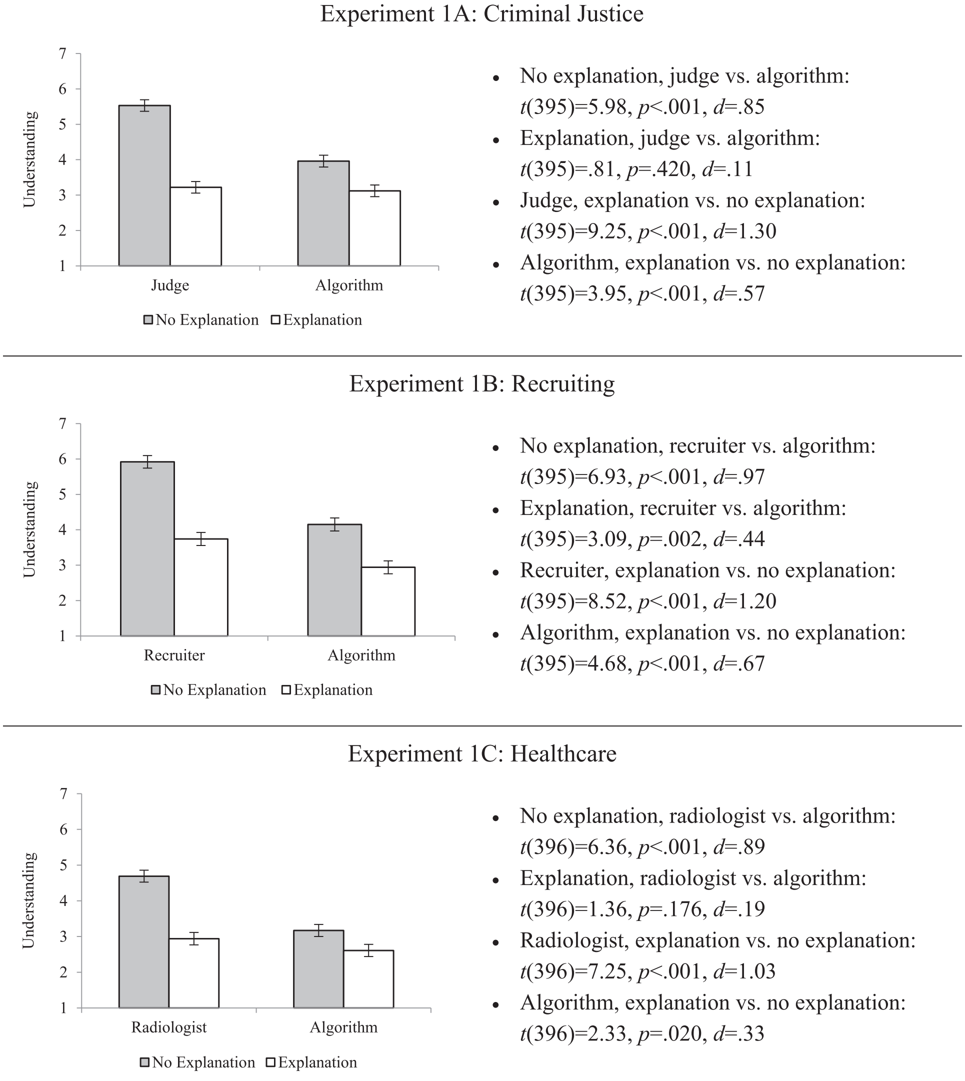
In each experiment, an analysis of variance (ANOVA) revealed the predicted interaction (1A: $F[1, 395] = 13.43$, $p < .001$, $\eta_p^2 = .03$; 1B: $F[1, 395] = 7.06$, $p = .008$, $\eta_p^2 = .02$; 1C: $F(1, 396) = 12.25$, $p < .001$, $\eta_p^2 = .03$). In the no-explanation conditions, respondents indicated a greater sense of understanding human than algorithmic decision-making. In the explanation conditions, this difference was reduced, as prompting respondents to explain the decision-making process lowered their sense of understanding more for the human than for the algorithm (see Figure 1 and Table 1). These results show that people foster a stronger illusion of understanding human than algorithmic decision-making.

Experiment 2

Experiment 2 tests the proposed projection mechanism by leveraging two paradigms used in prior research (Epley et al., 2009). First, by directly manipulating the egocentric reference point—that is, respondents' understanding of how they would make a decision. If people project their understanding of how they would make a decision more onto other humans than onto algorithms, changes in such egocentric reference point should have a stronger influence on judgments about another human than an algorithm.

Figure 1

Sense of Understanding as a Function of Decision-Maker and Explanation (±SE), and Planned Comparison Statistics in Experiments 1A–C



Experiment 1A: Criminal Justice

- No explanation, judge vs. algorithm: *t*(395)=5.98, *p*<.001, *d*=.85
- Explanation, judge vs. algorithm: *t*(395)=.81, *p*=.420, *d*=.11
- Judge, explanation vs. no explanation: *t*(395)=9.25, *p*<.001, *d*=1.30
- Algorithm, explanation vs. no explanation: *t*(395)=3.95, *p*<.001, *d*=.57

Experiment 1B: Recruiting

- No explanation, recruiter vs. algorithm: *t*(395)=6.93, *p*<.001, *d*=.97
- Explanation, recruiter vs. algorithm: *t*(395)=3.09, *p*=.002, *d*=.44
- Recruiter, explanation vs. no explanation: *t*(395)=8.52, *p*<.001, *d*=1.20
- Algorithm, explanation vs. no explanation: *t*(395)=4.68, *p*<.001, *d*=.67

Experiment 1C: Healthcare

- No explanation, radiologist vs. algorithm: *t*(396)=6.36, *p*<.001, *d*=.89
- Explanation, radiologist vs. algorithm: *t*(396)=1.36, *p*=.176, *d*=.19
- Radiologist, explanation vs. no explanation: *t*(396)=7.25, *p*<.001, *d*=1.03
- Algorithm, explanation vs. no explanation: *t*(396)=2.33, *p*=.020, *d*=.33

Second, by examining egocentric correlation—that is, the correlation between respondents' understanding of how they would make a decision and their understanding of how a human/algorithm makes the same decision. Projection implies that such a correlation should be stronger for a human than for an algorithm.

Method

Four hundred MTurk respondents were randomly assigned to a 2 (self-understanding: high, low) × 2 (decision-maker: human, algorithm) × 2 (understanding: preexplanation, postexplanation) mixed-design. Stimuli were adapted from Experiment 1A. Respondents read that parole decisions entail evaluating the risk that a defendant will reoffend if released. To manipulate respondents' sense of understanding how they would make this evaluation, we informed them about how easy or difficult it is for ordinary people to evaluate a defendant's risk to reoffend. In particular, respondents read that this evaluation is rather easy or very difficult to make and that ordinary people are pretty good or very bad at evaluating a defendant's risk to reoffend. We then asked respondents to indicate the extent to which they understood how they would evaluate the risk that a defendant will reoffend (1 = *do not understand at all* to 7 = *completely understand*). Then, they read that this evaluation is done by a judge/algorithm and rated their understanding of how a judge/algorithm evaluates a defendant's risk to reoffend on the same scale. All respondents were then prompted to explain the judge's/algorithm's decision-making process in detail before rerating their understanding of how a judge/algorithm evaluates a defendant's risk to reoffend (see online supplemental materials C).

Table 1
Descriptive Statistics in Experiments 1A–C

Experiment	Condition	*N*	Understanding *M* (*SD*)
1A	Judge	102	5.35 (1.31)
	Algorithm	98	3.96 (1.79)
	Judge explanation	101	3.22 (1.65)
	Algorithm explanation	99	3.03 (1.79)
1B	Recruiter	105	5.92 (1.46)
	Algorithm	98	4.15 (1.97)
	Recruiter explanation	97	3.74 (1.98)
	Algorithm explanation	99	2.94 (1.84)
1C	Radiologist	102	4.69 (1.67)
	Algorithm	101	3.17 (1.65)
	Radiologist explanation	97	2.94 (1.79)
	Algorithm explanation	100	2.61 (1.69)

Results

A 2 × 2 between-subjects ANOVA on self-understanding revealed only a main effect of the self-understanding manipulation, $F(1, 396) = 10.96$, $p = .001$, $\eta_p^2 = .03$; $M_{\text{High\_Self-Understanding}} = 4.53$, $SD_{\text{High\_Self-Understanding}} = 1.74$; $M_{\text{Low\_Self-Understanding}} = 3.98$, $SD_{\text{Low\_Self-Understanding}} = 1.61$, indicating that it was effective at changing respondent's sense of understanding how they would make the evaluation.

To test for projection, we examined whether the manipulation of self-understanding moderated the illusion of understanding human better than algorithmic decision-making. A 2 × 2 × 2 mixed-design ANOVA with repeated measures on understanding revealed the predicted three-way interaction, $F(1, 396) = 10.06$, $p = .002$, $\eta_p^2 = .03$ (Figure 2). To examine this interaction, we computed an IOED score by subtracting the understanding ratings for the judge/algorithm after the explanation task from the understanding ratings before the explanation task. In the high self-understanding conditions, respondents displayed a larger illusion of understanding for the judge ($M = 1.38$, $SD = 1.89$) than for the algorithm ($M = .72$, $SD = 1.75$), $t(396) = 2.65$, $p = .008$, $d = .38$, thus replicating Experiment 1A. In the low self-understanding conditions, no difference emerged ($M_{\text{judge}} = .48$, $SD_{\text{judge}} = 1.83$; $M_{\text{algorithm}} = .91$, $SD_{\text{algorithm}} = 1.41$), $t(396) = 1.83$, $p = .069$, $d = .25$. Manipulating respondents' own sense of understanding influenced the illusion of understanding the judge, $t(396) = 3.72$, $p < .001$, $d = .52$, but not the algorithm, $t(396) = .78$, $p = .437$, $d = .11$. See Figure 2 and Table 2, and online supplemental materials C for additional analyses.

As a further test of projection, we examined egocentric correlation—that is, the correlation between respondents' understanding of their own decision-making process and their preexplanation understanding of a judge's/algorithm's decision-making process. Respondents' understanding of how they would evaluate a defendant's risk was more strongly correlated with their understanding of how a judge ($r = .760$) than of how an algorithm would do so ($r = .478$, $z = 4.72$, $p < .001$). This pattern emerged in both the high ($r_{\text{judge}} = .796$, $r_{\text{algorithm}} = .494$; $z = 3.74$, $p < .001$) and low ($r_{\text{judge}} = .689$, $r_{\text{algorithm}} = .472$; $z = 2.36$, $p = .018$) self-understanding conditions.

Together, these results provide convergent evidence that the illusion of understanding human better than algorithmic decision-making emerges, at least in part, because people project their own intuitive understanding of a decision-making process more onto other humans than onto algorithms.

Experiment 3

Experiment 3 further tests the proposed projection mechanism by manipulating similarity. Prior research shows that projection is attenuated by asking people to elaborate on what makes them different from a target (Ames, 2004a, 2004b). If projection drives the illusion of understanding human better than algorithmic decision-making, prompting respondents to elaborate on dissimilarities between them and a decision-maker should reduce such illusion.

Figure 2
Sense of Understanding as a Function of Decision-Maker, Self-Understanding, and Explanation (±SE), in Experiment 2

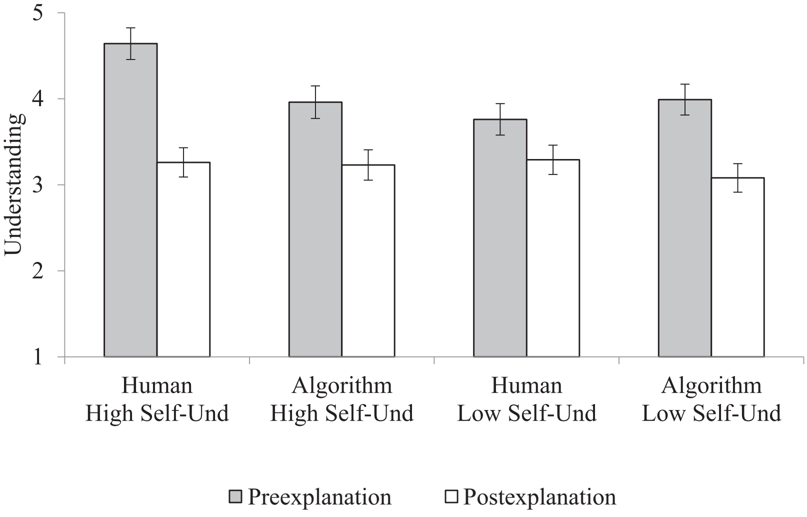


Table 2
Descriptive Statistics in Experiment 2

Condition	*N*	Understanding Preexplanation *M* (*SD*)	Understanding Postexplanation *M* (*SD*)
Human, high self-understanding	100	4.64 (1.90)	3.26 (1.74)
Algorithm, high self-understanding	94	3.96 (1.97)	3.23 (1.78)
Human, low self-understanding	101	3.76 (1.69)	3.29 (1.76)
Algorithm, low self-understanding	105	3.99 (1.78)	3.08 (1.54)

Method

Four hundred MTurk respondents were randomly assigned to a 2 (decision-maker: human, algorithm) × 2 (dissimilarity: control, dissimilar) × 2 (understanding: preexplanation, postexplanation) mixed-design. Stimuli were adapted from Experiment 1C. Respondents read that a radiologist/algorithm examines MRI images to diagnose osteoarthritis. In the control conditions, respondents rated their understanding of how a radiologist/algorithm examines MRI images to diagnose osteoarthritis. In the dissimilarity conditions, respondents were first prompted to elaborate on what made them different from a radiologist/algorithm, then rated their understanding (see online supplemental materials D for manipulation check). All respondents were then prompted to explain in detail how a radiologist/algorithm examines MRI images to diagnose osteoarthritis before rerating their understanding (see online supplemental materials D).

Results

A 2 × 2 × 2 mixed-design ANOVA with repeated measures on understanding revealed the predicted three-way interaction, $F(1, 396) = 10.13$, $p = .002$, $\eta_p^2 = .03$. In the control conditions, respondents displayed a significantly larger illusion of understanding—that is, difference in understanding pre- versus postexplanation—for the radiologist ($M = 1.06$, $SD = 1.54$) than for the algorithm ($M = .51$, $SD = 1.18$), $t(396) = 2.93$, $p = .004$, $d = .42$, thus replicating Experiment 1C. In the dissimilarity conditions, this was no longer the case ($M_{\text{radiologist}} = .36$, $SD_{\text{radiologist}} = 1.39$; $M_{\text{algorithm}} = .65$, $SD_{\text{algorithm}} = 1.15$), $t(396) = 1.57$, $p = .117$, $d = .22$. The dissimilarity manipulation significantly reduced the illusion of understanding the radiologist, $t(396) = 3.81$, $p < .001$, $d = .53$, but not the algorithm, $t(396) = .74$, $p = .463$, $d = .11$. See Figure 3 and Table 3, and online supplemental materials D for additional analyses. These results provide further evidence that projection drives, at least in part, the illusion of understanding human better than algorithmic decision-making.

An additional experiment provides further evidence for projection in the recruiting domain and shows that illusory understanding fosters greater trust in decisions made by humans than by algorithms (see online supplemental materials E).

General Discussion

Our work contributes to prior literature in two ways. First, it bridges two streams of research that have thus far been considered in isolation: IOED (Rozenblit & Keil, 2002) and projection (Krueger, 1998). IOED has mostly been documented for mechanical devices and natural phenomena and has been attributed to people confusing a superficial understanding of what something does for how it does it (Keil, 2003). Our research unveils a previously unexplored driver of IOED, namely, the tendency to project one's own cognitions onto

Figure 3
Sense of Understanding as a Function of Decision-Maker, Dissimilarity, and Explanation (±SE), in Experiment 3

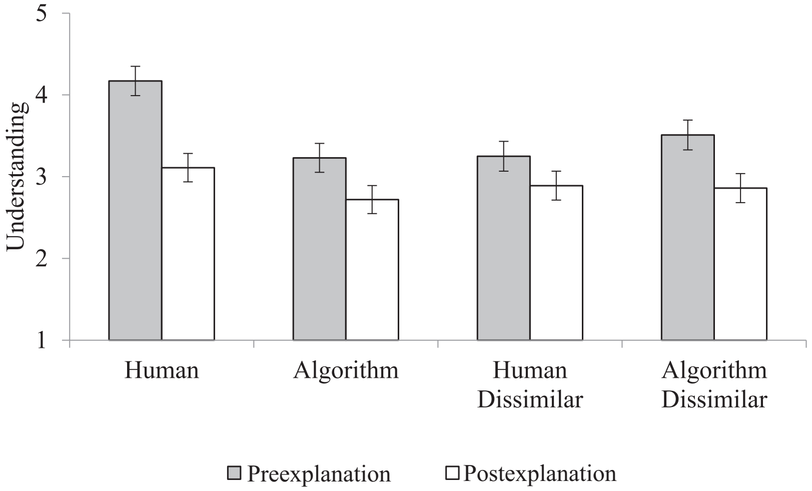


Table 3
Descriptive Statistics in Experiment 3

Condition	*N*	Understanding Preexplanation *M* (*SD*)	Understanding Postexplanation *M* (*SD*)
Human	101	4.17 (1.76)	3.11 (1.81)
Algorithm	98	3.23 (1.68)	2.72 (1.72)
Human dissimilar	104	3.25 (1.91)	2.89 (1.78)
Algorithm dissimilar	97	3.51 (1.84)	2.86 (1.70)

others, and in so doing extends the scope of IOED to human decision-making. Second, our work contributes to the literature on clinical versus statistical judgments (Meehl, 1954). Previous research shows that people tend to trust humans more than algorithms (Dietvorst et al., 2015). Among the many reasons for this phenomenon (see Grove & Meehl, 1996), one is that people do not understand how algorithms work (Yeomans et al., 2019). Our research suggests that people's distrust toward algorithms may stem not only from a lack of understanding how algorithms work but also from an illusion of understanding how their human counterparts operate.

Our work can be extended by exploring other consequences and psychological processes associated with the illusion of understanding humans better than algorithms. As for consequences, more research is needed to explore how illusory understanding affects trust in humans versus algorithms. Our work suggests that the illusion of understanding humans more than algorithms can yield greater trust in decisions made by humans. Yet, to the extent that such an illusion stems from a projection mechanism, it might also lead to favoring algorithms over humans, depending on the underlying introspections. Because people's introspections can be fraught with biases and idiosyncrasies they might not even be aware of (Nisbett & Wilson, 1977; Wilson, 2004), people might erroneously project these same biases and idiosyncrasies more onto other humans than onto algorithms and consequently trust those humans less than algorithms. To illustrate, one might expect a recruiter to favor people of the same gender or ethnic background just because one may be inclined to do so. In these circumstances, the illusion to understand humans better than algorithms might yield greater trust in algorithmic than human decisions (Bonezzi & Ostinelli, 2021).

As for psychological processes, our results provide evidence for projection as a mechanism that fosters the illusion of understanding humans better than algorithms. Yet, this phenomenon is likely multiply determined. Indeed, our results show that illusory understanding in part subsists when projection is disrupted. Future research is needed to identify other contributing processes. For instance, past research shows that an abstract construal level increases the IOED (Alter et al., 2010). It is possible that the tendency to construe humans more abstractly than algorithms (Kim & Duhachek, 2020) might also contribute to the phenomenon we document.

It is worth noting that the scope of our investigation is limited to situations in which people have no information about how a decision-maker—whether human or algorithmic—operates and can only draw on their subjective understanding of the decision-making process. Although this is typically the case in the domains that are the focus of our investigation, in other situations, people may have access to explanations for how the decision-making process works. In these cases, people's sense of understanding can be based on more objective information. This might either attenuate or further exacerbate the illusion of understanding humans better than algorithms, depending on the complexity of the algorithm.

Lastly, our findings raise a controversial question of societal relevance. Is algorithmic decision-making being held to unwarrantedly high transparency standards? Concerns about algorithms' black-box nature have sparked legislators to institutionalize the right to know how an algorithm reaches a determination (Goodman & Flaxman, 2017; Koene et al., 2019). The same right, however, is typically not invoked for human decision-makers. Yet, algorithms can often outperform human decision-makers (Dawes et al., 1989). Because the inner-workings of modern algorithms are often inexplicable (Castelvecchi, 2016; Goebel et al., 2018), holding inscrutable yet more accurate algorithms to transparency standards higher than those imposed on less accurate human counterparts that we delude ourselves to understand may ultimately be impractical and perhaps detrimental to societal welfare.

References

Allport, F. H. (1924). *Social psychology*. Houghton Mifflin.

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology*, *99*(3), 436–451. https://doi.org/10.1037/a0020218

Ames, D. R. (2004a). Inside the mind reader's tool kit: Projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology*, *87*(3), 340–353. https://doi.org/10.1037/0022-3514.87.3.340

Ames, D. R. (2004b). Strategies for social inference: A similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology*, *87*(5), 573–585. https://doi.org/10.1037/0022-3514.87.5.573

Ames, D. R. (2005). Everyday solutions to the problem of other minds: Which tools are used when? In B. F. Malle & S. D. Hodges (Eds.), *Other minds: How humans bridge the divide between self and others* (pp. 158–173). Guilford Press.

Ames, D. R., Weber, E. U., & Zou, X. (2012). Mind-reading in strategic interaction: The impact of perceived similarity on projection and stereotyping. *Organizational Behavior and Human Decision Processes*, *117*(1), 96–110. https://doi.org/10.1016/j.obhdp.2011.07.007

Artificial Intelligence Video Interview Act. (2020). *820 ILCS 42 §5*. http://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=4015&ChapterID=68

Bogen, M., & Rieke, A. (2018). *Help wanted: An exploration of hiring algorithms, equity, and bias. Technical report*. Upturn.

Bonezzi, A., & Ostinelli, M. (2021). Can algorithms legitimize discrimination? *Journal of Experimental Psychology: Applied*, *27*(2), 447–459. https://doi.org/10.1037/xap0000294

Bonezzi, A., Ostinelli, M., & Melzner, J. (2021). *The human black-box: The illusion of understanding humans better than algorithms*. https://doi.org/10.17605/OSF.IO/9TQEZ

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). *AI Now 2017 Report*. AI Now Institute at New York University.

Castelvecchi, D. (2016). Can we open the black box of AI? *NATNews*, *538*(7623), 20–23. https://doi.org/10.1038/538020a

Clement, R. W., & Krueger, J. (2002). Social categorization moderates social projection. *Journal of Experimental Social Psychology*, *38*(3), 219–231. https://doi.org/10.1006/jesp.2001.1503

Cohen, M. (2015). When judges have reasons not to give reasons: A comparative law approach. *Washington and Lee Law Review*, *72*(2), 483–571.

Davis, M. H. (2017). Social projection to liked and disliked targets: The role of perceived similarity. *Journal of Experimental Social Psychology*, *70*, 286–293. https://doi.org/10.1016/j.jesp.2016.11.012

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2016). Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychological Services*, *13*(3), 206–222. https://doi.org/10.1037/ser0000075

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. https://doi.org/10.1037/xge0000033

Epley, N., Converse, B. A., Delbosc, A., Monteleone, G. A., & Cacioppo, J. T. (2009). Believers' estimates of God's beliefs are more egocentric than estimates of other people's beliefs. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(51), 21533–21538. https://doi.org/10.1073/pnas.0908374106

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Estlund, C. (2010). Just the facts: The case for workplace transparency. *Stanford Law Review*, *63*, 351–407.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24*(6), 939–946. https://doi.org/10.1177/0956797612464058

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., & Holzinger, A. (2018). Explainable AI: The new 42? In A. Holzinger, P. Kieseberg, A. Tjoa, & E. Weippl (Eds.), *International Cross-domain Conference for Machine Learning and Knowledge extraction* (pp. 295–303). Springer.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation.". *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619. https://doi.org/10.1126/science.1134475

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Hayes, A. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*(8), 368–373. https://doi.org/10.1016/S1364-6613(03)00158-X

Keil, F. C., Rozenblit, L., & Mills, C. M. (2004). What lies beneath? Understanding the limits of understanding. In D. T. Levin (Ed.), *Thinking and seeing: Visual metacognition in adults and children* (pp. 227–249). MIT Press.

Kim, T. W., & Duhachek, A. (2020). The impact of artificial agents on persuasion: A construal level account. *Psychological Science*, *31*(4), 363–380. https://doi.org/10.1177/0956797620904985

Klehe, U., König, C. J., Richter, G. M., Kleinmann, M., & Melchers, K. G. (2008). Transparency in structured interviews: Consequences for construct and criterion-related validity. *Human Performance*, *21*(2), 107–137. https://doi.org/10.1080/08959280801917636

Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). *A governance framework for algorithmic accountability and transparency*. European Parliamentary Research Service.

Krueger, J. (1998). On the perception of social consensus. *Advances in Experimental Social Psychology*, *30*, 163–240. https://doi.org/10.1016/S0065-2601(08)60384-6

Krueger, J. (2007). From social projection to social behavior. *European Review of Social Psychology*, *18*(1), 1–35. https://doi.org/10.1080/10463280701284645

Krueger, J., & Zeiger, J. (1993). Social categorization and the truly false consensus effect. *Journal of Personality and Social Psychology*, *65*(4), 670–680. https://doi.org/10.1037/0022-3514.65.4.670

Krueger, J. (2000). The projective perception of the social world. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison. The Springer series in social clinical psychology* (pp. 323–351). Springer. https://doi.org/10.1007/978-1-4615-4237-7_16

Mangano, M. D., Bennett, S. E., Gunn, A. J., Sahani, D. V., & Choy, G. (2015). Creating a patient-centered radiology practice through the establishment of a diagnostic radiology consultation clinic. *American Journal of Roentgenology*, *205*(1), 95–99. https://doi.org/10.2214/AJR.14.14165

Mazurowski, M. A., Buda, M., Saha, A., & Bashir, M. R. (2017). Deep learning in radiology: An overview of the concepts and a survey of the state of the with focus on MRI. *Journal of Magnetic Resonance Imaging: JMRI*, *49*(4), 939–954. https://doi.org/10.1002/jmri.26534

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. https://doi.org/10.1037/11281-000

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology*, *87*(1), 1–32. https://doi.org/10.1016/j.jecp.2003.09.003

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *17*(8), 1306–1315. https://doi.org/10.1162/0898929055002418

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, *50*(4), 655–663. https://doi.org/10.1016/j.neuron.2006.03.040

Murray, B. (2012). Informed consent: What must a physician disclose to a patient? *The Virtual Mentor*, *14*(7), 563–566.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

O'Brien, E., & Ellsworth, P. C. (2012). More than skin deep: Visceral states are not projected onto dissimilar others. *Psychological Science*, *23*(4), 391–396. https://doi.org/10.1177/0956797611432179

Robbins, J. M., & Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review*, *9*(1), 32–47. https://doi.org/10.1207/s15327957pspr0901_3

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301. https://doi.org/10.1016/0022-1031(77)90049-X

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Smith, M. (2016, June 22). *In Wisconsin, a backlash against using data to foretell defendants' futures*. The New York Times. https://www.nytimes.com

Tamir, D. I., & Mitchell, J. P. (2013). Anchoring and adjustment during social inferences. *Journal of Experimental Psychology: General*, *142*(1), 151–162. https://doi.org/10.1037/a0028232

Van Boven, L., & Loewenstein, G. (2003). Social projection of transient drive states. *Personality and Social Psychology Bulletin*, *29*(9), 1159–1168. https://doi.org/10.1177/0146167203254597

Vigorita, M. S. (2003). Judicial risk assessment: The impact of risks, stakes, and jurisdiction. *Criminal Justice Policy Review*, *14*(3), 361–376. https://doi.org/10.1177/0887403403253722

Watson, H. J., & Nations, C. (2019). Addressing the growing need for algorithmic transparency. *Communications of the Association for Information Systems*, *45*(26), 488–510. https://doi.org/10.17705/1CAIS.04526

Wilson, T. D. (2004). *Strangers to ourselves*. Harvard University Press. https://doi.org/10.2307/j.ctvjghvsk

Woo, B. M., & Mitchell, J. P. (2020). Simulation: A strategy for mind-reading similar but not dissimilar others? *Journal of Experimental Social Psychology*, *90*, 104000. https://doi.org/10.1016/j.jesp.2020.104000

Yeomans, M., Shah, A. K., 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

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056

Appendix

Procedure and Stimuli for Experiments 1A–C

Method

Respondents recruited from MTurk were randomly assigned to a 2 (decision-maker: human, algorithm) × 2 (explanation: yes, no) between-subjects design. We aimed to collect 100 responses per condition. The sample size was predetermined and a sensitivity power analysis (Faul et al., 2009) indicated that the study had the power to detect a small effect ($\eta_p^2 = .02$) with a significance level α of .05 and a power (1-β) of .80. The final sample consisted of 400 respondents (208 females, 188 males, four did not indicate gender; age: *M* = 39.90, *SD* = 13.22) in Experiment 1A, 399 respondents (223 females, 173 males, three did not indicate gender; age: *M* = 39.37, *SD* = 13.23) in Experiment 1B, and 400 respondents (194 females, 201 males, five did not indicate gender; age: *M* = 39.58, *SD* = 13.29) in Experiment 1C, who passed an initial screening (see online supplemental materials B) and completed the survey.[1] There was no differential dropout across conditions (1A: $\chi^2(3) = .80$, $p = .849$; 1B: $\chi^2(3) = 1.62$, $p = .655$; 1C: $\chi^2(3) = 1.41$, $p = .703$).

Common Procedure

In all studies, participants read about a human [algorithmic] decision-maker. Specifically, respondents read about a judge [algorithm] evaluating a defendant's risk of recidivism (1A), a recruiter [algorithm] examining video interviews to evaluate applicants (1B), and a radiologist [algorithm] examining MRI images to diagnose a disease (1C). In the explanation condition, participants were first prompted to explain in detail the decision process of the respective human [algorithmic] decision-maker and then rated their understanding of the human's [algorithm's] decision-making process. In the no-explanation condition, respondents first rated their understanding of the human's [algorithm's] decision-making process and then were prompted to explain in detail the decision process of the respective human [algorithmic] decision-maker. Note that although the explanation task was irrelevant in the control conditions (because it was presented after the dependent variable), it ensured that all conditions were similar in terms of the effort required to complete the study; thus, reducing the risk of differential dropout (Zhou & Fishbach, 2016).

Stimuli for Experiment 1A

Respondents read that parole decisions entail an evaluation of a defendant's risk of reoffending made by a judge [an algorithm] (for a review on the use of risk assessment algorithms see Desmarais et al., 2016).

Decision-maker manipulation. In the United States, a criminal offender who has been sentenced to prison can become eligible for parole after serving part of the given sentence. When criminal offenders are paroled, they are released from prison and serve the remaining of the sentence in the community under supervision conditions. Decisions about parole entail an evaluation of the risk that a defendant will reoffend if released. In many jurisdictions, this evaluation is done by a judge who serves on a parole board [software called COMPAS that uses an algorithm] to evaluate a defendant's risk of recidivism.

Explanation manipulation: If you know it, please explain in detail the process used by the judge [the algorithm] to evaluate the risk that a defendant will reoffend if released. If there are aspects that you do not know or cannot explain, write "GAP" in your description at that point.

Measure of understanding: Do you understand how the judge [the algorithm] evaluates the risk that a defendant will reoffend if released? (1 = *do not understand at all* to 7 = *completely understand*).

Stimuli for Experiment 1B

Respondents read about an increasingly common practice whereby companies use recorded video interviews that are analyzed by a recruiter [algorithm] to screen job applicants (for a review on the use of algorithms in hiring see Bogen & Rieke, 2018).

[1] One participant failed to rate understanding; thus, only 399 data points are available for analysis.

(*Appendix continues*)

Decision-maker manipulation: Companies are turning to recorded video interviews to screen job applicants. Candidates answer questions in front of a camera, record a video, and send it to the employer. A recruiter [an algorithm] then reviews the video and evaluates the candidate.

Explanation manipulation: If you know it, please explain in detail the process used by a recruiter [an algorithm] to review a video and evaluate a candidate. If there are aspects that you do not know or cannot explain, write "GAP" in your description at that point.

Measure of understanding: Do you understand how a recruiter [an algorithm] reviews a video to evaluate a candidate? (1 = *do not understand at all* to 7 = *completely understand*).

Stimuli for Experiment 1C

Respondents read about the medical condition osteoarthritis and its diagnosis by means of MRI images, which are examined by a radiologist [an algorithm] (for a review on the use of algorithms in diagnostic imaging see Mazurowski et al., 2019).

Decision-maker manipulation: Osteoarthritis is a very common condition that affects millions of people worldwide. It occurs when the protective cartilage that cushions the ends of our bones wears down over time. It can affect any joint in the body, and it is most likely to affect the joints that we use most in everyday life, such as the joints of the hands, knees, feet, elbows, and neck. To diagnose osteoarthritis, a technician takes MRI images of the joints. The images are then examined by a radiologist [an artificial intelligence algorithm].

Explanation manipulation: If you know it, please explain in detail the process used by a radiologist [an artificial intelligence algorithm] to examine MRI images to diagnose osteoarthritis. If there are aspects that you do not know or cannot explain, write "GAP" in your description at that point.

Measure of understanding: Do you understand how a radiologist [an artificial intelligence algorithm] examines MRI images to diagnose osteoarthritis? (1 = *do not understand at all* to 7 = *completely understand*).

Received July 3, 2020
Revision received December 9, 2021
Accepted December 9, 2021 ■