Title
The computer judge: Expectations about algorithmic decision-making

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN
1069-7977

Authors
English, Sarah
Denison, Stephanie
Friedman, Ori

Publication Date
2021

Peer reviewed

The computer judge: Expectations about algorithmic decision-making

Sarah D English (senglish@uwaterloo.ca)
Department of Psychology, University of Waterloo

Stephanie Denison (stephanie.denison@uwaterloo.ca)
Department of Psychology, University of Waterloo

Ori Friedman (friedman@uwaterloo.ca)
Department of Psychology, University of Waterloo

Abstract

The use of algorithmic decision-making is steadily increasing, but people may have misgivings about machines making moral decisions. In two experiments (N = 551), we examined whether people expect machines to weigh information differently than humans in making moral decisions. We found that people expected that a computer judge would be more likely to convict than a human judge, and that both judge types would be more likely to convict based on individuating information than on base-rate information. While our main hypotheses were not supported, these findings suggest that people might anticipate machines will commit to decisions based on less evidence than a human would require, providing a possible explanation for why people are averse to machines making moral decisions.

Keywords: decision-making; base rates; Wells effect; theory of mind; algorithm aversion

Introduction

Artificial intelligence was formally introduced as a discipline in 1956 (Russell & Norvig, 1995). Since then, autonomous machines have been used to perform a wide variety of tasks like assisting with medical procedures (Parkin, 2016), calculating intricate flight patterns (Bartholomew-Biggs et al., 2003), and maintaining immensely complex inventory records (Cárdenas-Barrón et al., 2012). They also typically outperform humans on tasks that involve recall, strategy, and rationality (Chouard, 2016; Newborn, 2011; Markoff, 2011).

The extent to which algorithms are present in our daily lives is steadily increasing. However, despite the long-standing success of machines when it comes to decision-making, people are sometimes averse to machines making decisions. This tendency is referred to as algorithm aversion (Burton et al., 2019). Dietvorst and colleagues (2015) were the first to empirically investigate the concept. They asked participants to base their own incentivized predictions of graduate applicant success on the forecasts of either a human or an algorithmic model. Participants made these judgments after viewing feedback for each agent's performance, which included the agent's forecasting errors. Importantly, despite errors, the algorithmic model outperformed the human in all conditions. Seeing the algorithm err decreased participants' tendencies to bet on forecasts made by the algorithmic model, whereas seeing a human err did not decrease willingness to bet on the human forecasts. Furthermore, seeing the agents make forecasting errors decreased confidence in the accuracy of the algorithmic model, but not for the human. Together, these findings illustrate algorithm aversion, as participants demonstrated a reluctance to rely on algorithmic decisions, while maintaining a preference for the inferior human decision-maker.

While the findings of Dietvorst et al. (2015) provide strong support for algorithm aversion, there is some evidence that supports an opposing view. For instance, Awad and colleagues (2020) found that when a human and a machine both make an error that ultimately results in a fatal outcome, participants were actually *less* likely to attribute blame to the machine. This is discordant with findings of Dietvorst et al. (2015), which suggest that people are more willing to overlook a human error than the error of a machine, despite the errors being equivalent and resulting in the same (negative) outcome.

Recently, studies have suggested that people are averse to machines making decisions in moral situations due to the perceived lack of human mind in machines (Bigman & Gray, 2018). Mind perception research posits that people perceive minds along two dimensions: agency and experience (Gray et al., 2007). Similarly, it has been suggested that people also consider experience when determining decision-making abilities (Bigman & Gray, 2018); specifically, the ability to feel moral emotions (Malle & Scheutz, 2014), such as empathy, sympathy, and guilt (Tangney & Dearing, 2002). While people perceive human minds as possessing the components of agency and experience, machines inherently lack or merely approximate many of these components, (Gray et al., 2007; Gray & Wegner, 2012). As a result, the perception that machines have less agency and experience than humans may provide an explanation for why they are viewed as less fit to make morally charged decisions (Bigman & Gray, 2018).

In the present research, we explore a different question about how people view machine decision-making. We ask whether people anticipate that machines reach decisions differently than a human would, and whether people anticipate that humans and machines differ in the way they prioritize information when making decisions, including those which are morally relevant. If people expect machines to prioritize information differently than a human would, this might help explain why they are sometimes averse to machines making moral decisions (though we do not directly investigate this possibility).

Much previous work demonstrates that people often give more weight to indicant or individuating information than to base-rate information in judgment and decision-making tasks. That is, individuals often exhibit a tendency to ignore base-rate statistics in favor of information they perceive to be more pertinent to the situation, rather than integrating the two pieces of information (Bar-Hillel, 1980).

For the present work, a suitable example of this is one demonstrated by the classic cab problem (Bar-Hillel, 1980). Here, participants are presented with a scenario in which a taxicab is involved in a hit and run accident. It is noted that of all the cabs operating in the city, 85% are blue and only 15% are green. In court, a witness who was found to reliably identify each cab color 80% of the time, stated that a green cab caused the accident. Participants were then asked to indicate the probability that the cab causing the accident was green, as the witness claimed. To arrive at the correct answer in this scenario, one must accurately combine the various probabilities using Bayesian analyses to conclude that there is approximately a 41% chance that a green cab caused the accident. Despite this, most participants provided estimates of over 50%, with some indicating upwards of an 80% probability that the at-fault cab was green. This example further highlights reliance on individuating information over base-rate statistics. As a result, previous work pertaining to probability interpretation has been concerned with the issue of people deriving subjective probabilities that do not align with the mathematically correct answer (see Kahneman, 2011). While having a robust understanding of probability is beneficial in many aspects of day-to-day life, there are certain contexts in which appropriately interpreting probability is exceptionally important. For example, decisions based on probabilistic information in legal situations may have particularly consequential outcomes.

What evidence do we base our legal decisions on? Is a guilty man rightfully convicted, or does he walk free? Is an innocent man acquitted, or wrongfully convicted? How can we be confident that the right decision was made? To answer the first question, legal scholars have argued that people are resistant to naked statistical evidence when assigning guilty verdicts; where "naked statistics" refer to probabilities that are not case specific, as the evidence did not result from the event in question, but rather existed prior to, or independently of the specific case (Wells, 1992).

Wells conducted a series of experiments to explore the impact of naked statistical evidence on people's decisions about verdicts. In contrast to previous studies, he presented participants with probability information that would be intuitively processed, wherein the subjective probability and mathematically correct answer would be in agreement. Wells (1992) presented participants with a description of a legal case in which a woman is suing the Blue Bus Company for causing the death of her dog. It was explained that while she saw a bus hit her dog, she could not determine its color. In the standard version of the task, participants received information regarding the distribution of buses in the area where the dog was hit: the Blue Bus Company was said to own 80% of all buses, while the Grey Bus Company owned 20%. Across five studies, Wells presented versions of the accident description, which featured information regarding eye-witness testimony, causal relevance, as well as case-specific statistical information, respectively. After reading the description, participants either indicated whether they would convict the Blue Bus Company, or estimated the probability that a blue bus hit the dog.

As expected, Wells found that participants' subjective probability judgments approximated the statistically correct probability (80%) in all versions of the case. However, profound differences emerged in verdict decisions. Despite the information being statistically equivalent, participants only returned guilty verdict decisions that were comparable to the subjective probability ratings when the evidence was perceived as being more directly connected to the specific case (e.g., testimony from a witness working on the day of the accident), rather than when information existed independently of the case (e.g., the number of buses owned by each company).

In the current research, we wondered whether a greater preference for individuating information over base-rate information might contribute to people's general distrust of machines making moral decisions. Presumably, people expect other humans to share their tendency to base decisions on individuating information. However, they might expect computers to instead favor base-rate information. This prediction might reflect a stereotype that computers make decisions based on purely statistical considerations, without insight into details pertaining to the specific case at hand. Given this, a computer faced with Wells' (1992) task might be more likely to return a guilty verdict when presented with information regarding the distribution of buses or the rate of accidents in the area, in comparison to receiving witness testimony information. On the other hand, if the computer considers both types of information to be equivalent, given that the statistical information is identical (85%) for both evidence types, it may be equally likely to return a guilty verdict in both instances.

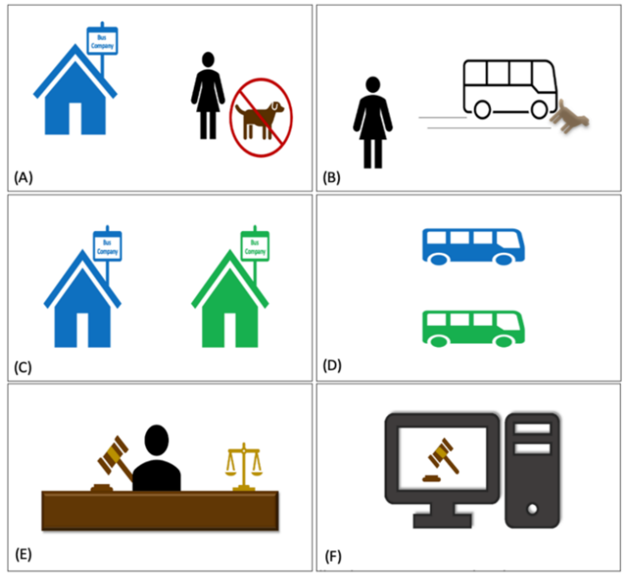
Experiment 1

Experiment 1 investigated whether people's judgments about the decisions different agents would make in a hypothetical legal situation would differ depending on the type of evidence that was presented. This research received ethics clearance through the University of Waterloo. See the pre-registration here: https://osf.io/85taj/?view_only=63d2ef05d3654c44bec74ba72a2b1085.

Method

Participants Participants were 234 adults from the United States (M_{age} = 41.27, *SD* = 12.37; range = 22 – 83; 39% female) who completed a survey on CloudResearch for 50 cents. An additional 17 participants completed the survey but were excluded from the analysis due to incorrectly answering the comprehension questions, or for failing to answer any of the test questions.

Materials and procedure Participants were randomly assigned to one of two conditions. In each condition, participants read a scenario in which the Blue Bus Company is on trial for causing the death of a woman's dog (see Figure 1 for a sample script). In the Base Rate Condition, the evidence stated that "85% of buses in the area are owned by the Blue Bus Company. 15% of buses are owned by the Green Bus Company". In the Witness Testimony Condition, the evidence stated that "85% of witnesses indicated that the bus that caused the accident was blue. 15% of witnesses indicated that the bus that caused the accident was green". Participants were then shown two types of judges: Judge Brown, a human judge, and JudgeComp, an autonomous computer system (presented in random order). Using a 7-point Likert scale (from *Definitely No* to *Definitely Yes*), participants were asked to indicate whether Judge Brown and JudgeComp would convict the Blue Bus Company, based on the evidence that was presented.



(A) The Blue Bus Company is on trial for killing a woman's dog. (B) The woman saw a bus negligently run her dog over. But she is colorblind, so she could not see the color of the bus. (C) However, the Blue Bus Company is suspected because: Only the Blue Bus Company and the Green Bus Company operate in the area where the dog was hit, and … (D) 85% of buses operating in the area are blue. 15% of buses operating in the area are green. (E) Imagine that Judge Brown is responsible for determining the verdict. *Will Judge Brown convict the Blue Bus Company for killing the dog?* (F) Imagine that JudgeComp is responsible for determining the verdict. *Will JudgeComp convict the Blue Bus Company for killing the dog?*

Figure 1: Script and slides from the Base Rate Condition in Experiment 1.

Results and Discussion

Of primary interest was whether individuals would give different conviction ratings depending on the type of judge presented, and whether these ratings would differ depending on evidence type. Figure 2 shows the mean conviction ratings for each judge type based on evidence type. A Generalized Estimating Equations (GEE) ordinal logistic regression with evidence type as a between-subjects factor (bus distribution evidence, witness testimony evidence), judge type as a within-subjects factor (Judge Brown, JudgeComp), and their interaction, revealed a significant main effect of evidence type, *Wald* X^2 (df = 1, N = 234) = 36.64, $p < .001$, and a significant main effect of judge type, *Wald* X^2 (df = 1, N = 234) = 45.65, $p < .001$. There was no significant interaction (p = .314).

These findings suggest that in each condition, participants believed that JudgeComp, the autonomous computer system, would be more likely to convict the Blue Bus Company than the human judge. Moreover, they rated both Judge Brown and JudgeComp as more likely to convict based on witness testimony as evidence than based on the distribution of buses. However, against our expectations, there was no interaction. This finding conflicted with our prediction that participants would expect the computer judge to give relatively more weight to the base-rate information, compared to the human judge. One possible explanation is that participants expected the computer judge to place more weight on the witness evidence due to the fact that there were *many* witnesses.

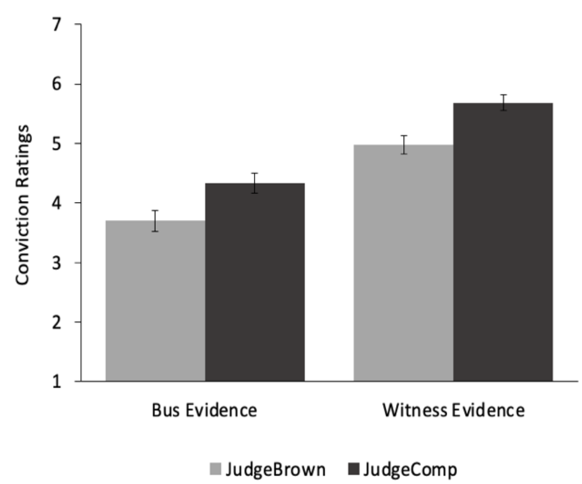


Figure 2: Mean conviction ratings for each judge type based on evidence type condition. Error bars represent ± 1 standard error.

Thus, in a sense, the witness testimony condition also provided a distribution (i.e., the sample of witnesses). This differed from the original Wells' (1992) experiment. Therefore, in the next experiment, we adopted from Wells' original study by including only one witness in the witness testimony condition.

We were also intrigued by how often participants predicted that the human judge would convict based on base-rate evidence. In the original Wells experiments, participants overwhelmingly denied that they would convict the Blue Bus Company based on this evidence. One key difference between the original procedure and that in Experiment 1 is that we only asked participants to predict whether *others* would be in favor of convicting. In the next study, we addressed this by including an additional condition in which participants were asked to make first-person judgments about the likelihood that *they* would convict.

In an additional effort to replicate Wells' original findings, we also presented a revised test question, such that it was more similar to the original test question. In Experiment 2, we specified that the convicted bus company would be forced to pay damages, as the original experiment indicated that the bus company was being sued (see Wells, 1992).

Experiment 2

Similar to the previous experiment, Experiment 2 also investigated whether people's judgments about the decisions various agents would make in a hypothetical legal situation would differ depending on the type of evidence that was presented. See the pre-registration here: https://osf.io/85taj/?view_only=63d2ef05d3654c44bec74ba72a2b1085.

Method

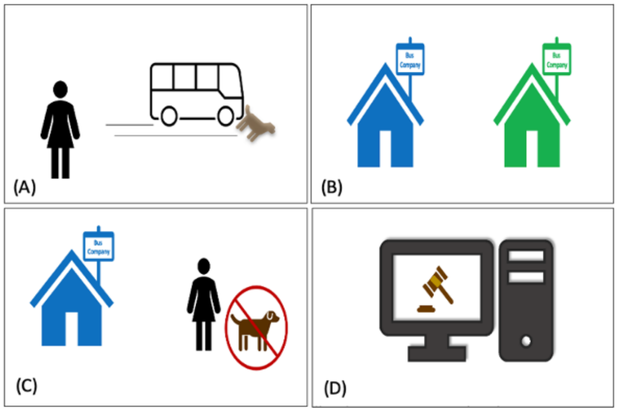
Participants Participants were 317 adults from the United States (M_{age} = 40.16, *SD* = 12.07; range = 18 – 77; 49% female) who completed a survey on CloudResearch for 50 cents. An additional 34 participants completed the survey but were excluded from the analysis due to incorrectly answering the comprehension questions, or for failing to answer any of the test questions.

Materials and procedure Participants were randomly assigned to one of three conditions: a human judge condition ("Judge Brown"), an autonomous computer system condition ("JudgeComp"), and a condition in which the participant was told that they were the judge. In each condition, participants read a scenario in which the Blue Bus Company is on trial for causing the death of a woman's dog (see Figure 2 for a sample script). Participants were then presented with two types of evidence (presented in random order) suggesting the Blue Bus Company was responsible for the accident. One piece of evidence stated that "85% of buses in the area are owned by the Blue Bus Company, and only 15% are owned by the Green Bus Company". The other piece of evidence stated that "A man who witnessed the accident said that it was caused by a blue bus. At night, he accurately identifies bus colors 85% of the time. He is inaccurate 15% of the time." After reading about each piece of evidence, participants used a 7-point Likert scale (from *Extremely Unlikely* to *Extremely Likely*), to rate the likelihood that the judge would convict the Blue Bus Company and force them to pay damages. Note that the wording of the question was changed across between-subjects' conditions to reflect judge-type. For example, in the JudgeComp condition, participants were asked *"How likely is JudgeComp to convict the Blue Bus Company and force them to pay damages?"*

Results and Discussion

Of primary interest was whether individuals would give different conviction ratings depending on the type of evidence presented, and whether these ratings would differ depending on judge type. Figure 4 shows the mean conviction ratings for each evidence type based on judge type. A GEE ordinal logistic regression with judge type as a between-subjects factor (Judge Brown, JudgeComp, You), evidence type as a within-subjects factor (bus distribution evidence, witness testimony evidence), and their interaction, revealed a significant main effect of judge type, *Wald* X^2 (df = 2, N = 317) = 19.42, $p < .001$, and a significant main effect of evidence type, *Wald* X^2 (df = 1, N = 317) = 18.77, $p < .001$. There was no significant interaction (p = .864). Pairwise comparisons (which required the data to be treated as linear) revealed that ratings for JudgeComp (M = 5.30, SE = 0.14) significantly differed from those for Judge Brown (M = 4.61, SE = 0.15; $p < .001$), and from the condition in which the participant was the judge ("You"; M = 4.36, SE = 0.18; $p < .001$). There was no significant difference between Judge Brown and the "You" condition (p = .287).

These findings suggest that in each condition, participants believed that they, and the fictional judges, would be more likely to convict based on witness testimony than based on the distribution of buses. Additionally, they rated JudgeComp as most likely to convict based on both types of evidence. As in the previous experiment, there was no interaction between judge and evidence type. Also, as before, we did not see any strong denials in any condition that convictions should be based on base-rate evidence. We also failed to find strong denials when participants made first-person judgments when indicating the likelihood that they would convict the Blue Bus Company.



(A) One night, a bus ran over a woman's dog.

(B) Only the Blue Bus Company and the Green Bus Company operate in the area where the dog was hit. However, the dog's owner is colorblind, so she could not see the color of the bus.

(C) Right now, the Blue Bus Company is on trial for killing the woman's dog.

(D) JudgeComp is responsible for determining the verdict.

Base rate evidence and question: Suppose this is the evidence suggesting the Blue Bus company is responsible: **85% of buses in the area are owned by the Blue Bus Company, and only 15% are owned by the Green Bus Company.** *How likely is JudgeComp to convict the Blue Bus Company and force them to pay damages?*

Witness testimony evidence and question: Suppose this is the evidence suggesting the Blue Bus Company is responsible: **A man who witnessed the accident said that it was caused by a blue bus. At night, he accurately identifies bus colors 85% of the time. He is inaccurate 15% of the time.** *How likely is JudgeComp to convict the Blue Bus Company and force them to pay damages?*

Figure 3: Script and slides from the JudgeComp condition in Experiment 2. Note that the order in which each type of evidence was displayed was counterbalanced across participants.

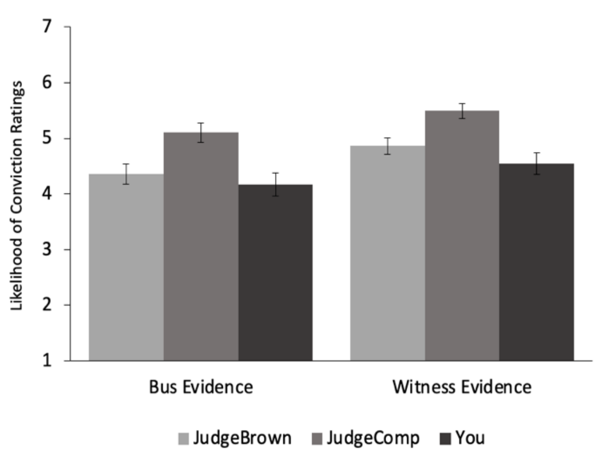


Figure 4: Mean conviction ratings for each evidence type based on judge type condition. Error bars represent ± 1 standard error.

General Discussion

In two experiments, we examined whether participants would have different intuitions about the conviction decisions that either a human or a computer judge would make based on the evidence type (i.e., base-rate or individuating information) that was presented during the trial. To our knowledge, our studies are among the first to explore whether people think the way algorithms make decisions differs from the way humans make decisions (see also Longoni et al., 2019, for related work).

Our findings show that people expect that an autonomous computer system will be more likely to convict a defendant than a human judge, and that both agents would be more willing to convict when presented with individuating information in the form of witness testimony, than when presented with base-rate information. Contrary to our expectations, we did not find an interaction between judge and evidence type. We hypothesized that participants might rate the computer as more willing to convict on the basis of base-rate information and a human judge more likely to convict on the basis of individuating information.

However, participants expected the computer judge to convict more often overall. Although participants did not expect the computer judge to weigh the two types of evidence *differently*, they did expect it to weigh all evidence more *heavily* than a human judge. While we did not anticipate this finding, it may provide insight into why people are averse to machines making moral judgments. Specifically, our findings suggest that people might anticipate that machines will commit to decisions based on less evidence than a human would require. As with many moral decisions, such hair-trigger commitment could have worrisome consequences. For example, in a legal context, it may lead to rash decisions resulting in false convictions.

However, another possibility is that people might expect that machines will be more likely to reach decisions based on less information than humans, irrespective of outcome. That is, people might expect a computer judge to be equally as likely to give credit and rewards as it would punishments, in each case based on less information than a human judge would require. On this view, if we had asked participants how likely the computer judge is to make a positive or non-punitive judgment, they would have also expected it to make a more sensitive judgment than a human judge. Further research will be needed to explore this possibility.

One perplexing, and rather unexpected finding, was the high rating of affirmative conviction based on base-rate information for both the human and computer judge. This result is inconsistent with the findings of Wells (1992), which revealed strong denials for willingness to convict on base-rate information alone. One possible explanation for the discrepancies between our findings and those of Wells could be the difference in how the dependent variable was measured. In both of our studies, participants were asked to rate the likelihood of conviction using a 7-point Likert scale, while Wells obtained conviction ratings using a binary, forced-choice rating scale. Further, it is possible that we would have seen a greater sensitivity to evidence type if the conviction decision resulted in a more personal consequence. That is, if an affirmative conviction would result in the imprisonment of the individual driver of the bus involved in the accident.

Acknowledgements

This work was funded by grants to O.F. and S.D. from the Natural Sciences and Engineering Research Council of Canada and from the Social Sciences and Humanities Research Council of Canada.

References

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour, 4*(2), 134–143.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44,* 211–233.

Bartholomew-Biggs, M. C., Parkhurst, S. C., & Wilson, S. P. (2003). Global optimization approaches to an aircraft routing problem. *European Journal of Operational Research, 146*(2), 417–431.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181,* 21–34.

Burton, J. W., Stein, M., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making, 33*(2), 220–239.

Cárdenas-Barrón, L. E., Treviño-Garza., & Wee, H. M. (2012). A simple and better algorithm to solve the vendor managed inventory control system of multi-product multi-constraint economic order quantity model. *Expert Systems with Applications, 39*(3), 3888–3895.

Chouard, T. (2016). The Go Files: AI computer clinches victory against Go champion. *Nature*.

Clarke, R. (1992). Free will and the conditions of moral responsibility. *Philosophical Studies, 66*(1), 53–72.

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.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315,* 619.

Gray, K., Young, L., & Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry, 23*(2), 101–124.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research, 46*(4), 629–650.

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, ETHICS 2014.*

Markoff, J. (2011). On 'Jeopardy!' Watson Win Is All but Trivial. Retrieved January 28, 2021, from https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html

Mele, A., & Sverdlik, S. (1996). Intention, intentional action, and moral responsibility. *Philosophical Studies, 82*(3), 265–287.

Newborn, M. (2011). *Beyond Deep Blue: Chess in the Stratosphere*. New York, NY: Springer.

Parkin, S. (2016). The Artificially Intelligent Doctor Will Hear You Now – MIT Technology Review. Retrieved January 28, 2021, from https://www.technologyreview.com/s/600868/the-artificially-intelligent-doctor-will-hear-you-now/

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach.* Englewood Cliffs, NJ: Prentice-Hall.

Tangney, J. P., & Dearing, R. (2002). Shame and Guilt. New York, NY: Guilford Press.

Wells, G. L. (1992). Naked statistical evidence of liability: Is subjective probability enough? *Journal of Personality and Social Psychology, 62,* 739–752.