Judgment and Decision Making, Vol. 17, No. 3, May 2022, pp. 628–645

Expectations of how machines use individuating information and base-rates

Sarah D. English* Stephanie Denison† Ori Friedman‡

Abstract

Machines are increasingly used to make decisions. We investigated people's beliefs about how they do so. In six experiments, participants (total $N = 2664$) predicted how computer and human judges would decide legal cases on the basis of limited evidence — either individuating information from witness testimony or base-rate information. In Experiments 1 to 4, participants predicted that computer judges would be more likely than human ones to reach a guilty verdict, regardless of which kind of evidence was available. Besides asking about punishment, Experiment 5 also included conditions where the judge had to decide whether to reward suspected helpful behavior. Participants again predicted that computer judges would be more likely than human judges to decide based on the available evidence, but also predicted that computer judges would be relatively more punitive than human ones. Also, whereas participants predicted the human judge would give more weight to individuating than base-rate evidence, they expected the computer judge to be insensitive to the distinction between these kinds of evidence. Finally, Experiment 6 replicated the finding that people expect greater sensitivity to the distinction between individuating and base-rate information from humans than computers, but found that the use of cartoon images, as in the first four studies, prevented this effect. Overall, the findings suggest people expect machines to differ from humans in how they weigh different kinds of information when deciding.

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

*Department of Psychology, University of Waterloo. https://orcid.org/0000-0003-4269-1121.

†Department of Psychology, University of Waterloo. http://orcid.org/0000-0002-6658-4139.

‡Corresponding author. Department of Psychology, University of Waterloo, 200 University Avenue W, Waterloo, Ontario, Canada N2L 3G1. Email: friedman@uwaterloo.ca. https://orcid.org/0000-0003-2346-9787.

Research was supported by separate Social Sciences and Humanities Research Council of Canada awarded to SD and to OF.

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

1 Introduction

Autonomous computers perform a wide variety of tasks and are increasingly used in decision making. Here we investigate people's beliefs about how autonomous computers make decisions. Specifically, we ask whether people expect computers and humans to differ in how they prioritize information when deciding, and whether people expect computers to differ from humans in the decisions they reach. This work brings judgment and decision making research on the distinction between individuating information and base-rates into contact with research on theory of mind and social cognition more generally. Also, this work contributes to knowledge of people's "theory of machine" — their lay beliefs about the internal aspects of machine decision making (Logg, 2021; Logg et al., 2019).

Some recent work is broadly consistent with the possibility that people expect machines to differ from humans in how they prioritize information when making decisions. For example, children and adults expect robots to make different choices than humans when faced with the same situation (Flanagan et al., 2021). Also, people often rate computers and robots as less able than humans to have many mental states, but especially desires and emotions (Bigman & Gray, 2018; Gray et al., 2007; Haslam et al., 2008; Weisman et al., 2017). These findings suggest that people might expect computers to be oblivious to emotional considerations when reaching decisions. Consistent with this, people do expect computers to struggle with tasks that have a subjective element, and which depend on emotions and intuitions (Castelo et al., 2019). For example, whereas people are similarly willing to take advice from a human or computerized advisor for financial decisions (viewed as depending on objective facts), they greatly prefer receiving advice from humans for decisions viewed as having a subjective element, such as those pertaining to dating.

People also expect machines to exhibit uniqueness neglect — a tendency to treat situations in the same way, and to overlook distinctive circumstances (Longoni et al., 2019). Consistent with this, people rate computerized medical practitioners as less able than human ones to recognize the uniqueness of their medical conditions. People are also reluctant to recommend computers as medical practitioners for others when those individuals are described as having unique medical conditions. Similarly, people may view machines as comparatively inflexible — they might believe that computers are capable of doing only what they have been programmed to do (Laakasuo et al., 2021; also see Kim & Duhachek 2020).

1.1 Two kinds of information

We investigate whether people expect humans and machines to differ in how they use two kinds of information. Following classic work on judgment and decision making, we contrast individuating information and base-rates (Kahneman & Tversky, 1973). As we will review, people often make decisions by prioritizing individuating information, while neglecting and underweighting base-rate information. We hypothesize that people likely expect their

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

fellow humans to make decisions in this way. But people might not expect the same from machines such as autonomous computers. For example, they might expect machines to give similar weight to both kinds of information.

The priority people give to individuating information over base-rates is seen in the classic cab problem. Participants learn about a taxicab involved in a hit-and-run (Bar-Hillel, 1980; Tversky & Kahneman, 1977). The accident occurs in a city where 85% of the cabs are blue and 15% are green. But a witness says the responsible cab is green, and this witness reliably identifies the cabs by color 80% of the time. In this example, the witness testimony is individuating information — it is specific to the cab that caused the accident. In contrast, the information about the proportion of blue and green cabs is base-rate information — rather than concerning a specific cab, it pertains to the distribution of cabs in the city. This distinction is closely related to that between "inside" versus "outside" information (e.g., Kahneman & Tversky, 1982; Lagnado & Sloman, 2004). Correctly determining the liklelihood that the accident was caused by a green cab (41%) requires considering both kinds of information. Specifically, it requires integrating the information via Bayes rule. But people often neglect the base-rate information and instead focus on the testimony from the witness.

People also prioritize individuating information over base-rate information in simpler judgments where there is no need to integrate the two kinds of information and no need to engage in complex math. For example, in research on legal decision making, participants read about a bus company on trial for causing the death of woman's dog (Wells, 1992). Each participant had access to just one kind of evidence. For example, they either received base-rate information to the effect that 80% of the buses in the area were operated by the Blue Company (rather than a rival company), or they received individuating evidence that witness testimony of 80% accuracy implicated a bus from the Blue company (individuating information). Here, the type of evidence did not affect participants' judgments about the *likelihood* that the dog had been killed by a blue bus. But evidence type did affect recommendation about the court case *verdict*. Participants were much more likely to recommend a guilty verdict when the Blue Bus company was implicated based on individuating information than on base-rates (for replications and extensions of this work, see Arkes et al., 2012; Niedermeier et al., 1999; Turri et al., 2017).

Such findings suggest people give greater evidential weight to individuating information than base-rate information in legal decisions (also see Koehler, 2001). However, people might expect computers to be insensitive to the difference between these kinds of information, and to treat the two kinds of information similarly. This prediction might reflect beliefs that computers make decisions based on purely numerical and statistical considerations, without insight into what the numbers and statistics actually mean (Searle, 1980; Weisman et al., 2017). In the example of the bus that runs over a dog, a computer might be expected to treat statistical evidence equivalently regardless of whether this evidence is specifically about the bus that ran over the dog (individuating information) or about the

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

distribution of buses in the city (base-rate information).

1.2 The current experiments

Across six experiments, we investigated the hypothesis that people expect computers to be less sensitive than humans to the distinction between individuating and base-rate information. In our first four experiments, participants read different versions of the court case stemming from an accident where a bus runs over a woman's dog. Whereas the focus of earlier work was on participants' own recommendations of whether one bus company should be found guilty (e.g., Wells, 1992), our studies mainly focused on their predictions of others' judgments. Specifically, we asked participants to predict whether a human judge and a machine judge would reach this verdict. Following Bigman and Gray (2018), we probed judgments about machines by asking them to consider the workings of an autonomous computer.

The findings from the first four experiments prompted us to consider further hypotheses we had not considered at the outset: (1) People might expect machines to initiate punishment on the basis of less evidence than a human would require; (2) people might expect machines to also make other kinds of decisions based on less evidence than a human would require; and (3) people's predictions of how computers will weigh evidence might depend on how scenarios about computers are conveyed. Our final experiments examined these hypotheses, while using somewhat different methods than the earlier experiments.

1.3 General methods

Preregistrations, data, and code for all experiments are available at https://osf.io/85taj/. In each experiment, participants were residents of the United States, and tested using Qualtrics and CloudResearch. We used the "block low quality participants" option and required participants to have a HIT approval rate of 95–100% over at least 100 prior HITS. In each experiment, we also blocked individuals who had completed any prior experiment in the series. After completing the main task in each experiment, participants answered multiple-choice attention checks and questions about their age and gender. We excluded participants who failed any of the attention check questions and participants who neglected to respond to any test questions. We analyzed results from all experiments except Experiment 4 using ANOVAs. In Experiment 4, results were analyzed using a generalized estimating equations model for binary logistic data.[1]

[1]Our preregistratrations specified running generalized estimating equation models for all experiments. However, switching to ANOVAs did not change any major results.

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

2 Experiments 1 to 4

In each of these experiments, participants read about a court case where the Blue Bus Company is on trial for causing the death of a woman's dog, and then predicted whether the bus company would be found guilty. In each experiment, we manipulated two factors, evidence type and judge type. The evidence against the bus company came either from witness testimony or from base-rates concerning the percentage of Blue and Green buses operating in the area. The judge deciding the case was either a regular human judge or an autonomous computer system, though one experiment also included a condition where participants responded as though they were the judge deciding the case.

We report these experiments together because they all yielded the same pattern of findings. Participants in each experiment were sensitive to the difference between individuating information and base-rates, but the effect was never as pronounced as in Wells' (1992) original research. Most methodological changes across these experiments were adopted to see if we could come closer to replicating the original patterns[2]; see Table 1 for a summary of the design of each experiment. For example, whereas our experiments primarily asked participants to predict whether a judge would find the Blue company guilty, Experiment 2 included a condition where participants gave their own judgment of whether they themselves would convict the company, much as Wells (1992) had asked participants for their recommendations.

Table 1: Summary of the experimental designs of Experiments 1 to 4.

Experiment	Evidence	Judge	Rates	Measure
1	between	within	85% vs 15%	7-point Likert
2	within	between	85% vs 15%	7-point Likert
3	within	between	98% vs 2%	7-point Likert
4	within	between	80 % vs 20%	yes-or-no

[2]In retrospect, the differences in findings are unsurprising. One reason is that our experiments differed from those of Wells (1992) in many ways. For example, our scenarios were much shorter. Another reason is that other experiments that did use the original stories reported smaller differences between conditions than the original paper (e.g., Arkes et al., 2012; Niedermeier et al., 1999). We should also mention that the diminished effect of type-of-evidence that we observed could not just have resulted from order effects and our use of within-subject manipulations. Consider our fourth experiment, which asked for binary judgments (as did the original experiments). When restricting the results to the first type of evidence each participant saw, 62% of participants thought the human judge would punish based on witness testimony and 44% thought the human judge would convict with base-rates. By comparison, the Experiment 1 in Wells 1992) had corresponding rates of 67% and 8%.

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

2.1 Method

2.1.1 Participants

In Experiment 1, we tested 234 participants (M_{age} = 41.27, SD = 12.37; range = 22–83; 39% female); an additional 17 were excluded. In Experiment 2, we tested 317 participants (M_{age} = 40.16, SD = 12.07; range = 18–77; 49% female); an additional 34 were excluded. In Experiment 3, we tested 236 participants (M_{age} = 39.89, SD = 12.51; range = 18–78; 44% female); an additional 17 participants were excluded. Finally, in Experiment 4, we tested 235 participants (M_{age} = 41.62, SD = 13.59; range = 20–79; 48% female); an additional 15 were excluded. In these experiments, we attempted to recruit enough participants so that we would have about 100 per between subjects condition after exclusions; the sample size in Experiment 2 was larger than in the other experiments because it had three between subjects conditions rather than two.

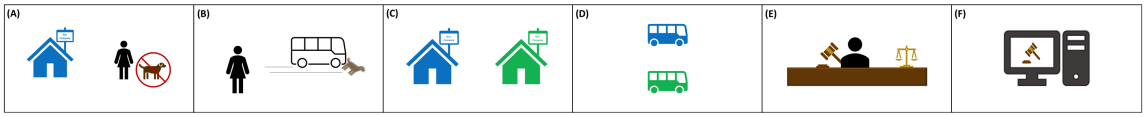
2.1.2 Procedure

In Experiment 1, evidence type was manipulated between subjects; see Figure 1 for sample stimuli. In the base-rate condition, 85% of buses were owned by the Blue Bus Company, and 15% were owned by the Green Bus Company; in the witness condition, 85% of witnesses said the accident was caused by a blue bus, and 15% said it was caused by a green one. Judge type was manipulated within subjects with each participant predicting whether the computer and human judges would convict the company. The test question for each judge was presented on a separate screen and presentation order was random. Participants indicated predictions using a 7-point Likert scale ranging from "Definitely No" (1) to "Definitely Yes" (7).

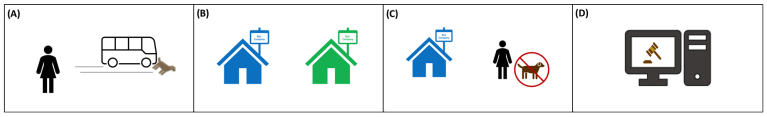
In Experiment 2, judge type was manipulated between subjects. After first reading about the case, participants either learned it would be decided by a regular human judge, an autonomous computer system, or the participant themselves. Evidence type was manipulated within subjects. Across separate screens (presented in random order), participants were asked to imagine two hypothetical situations differing in the kind of evidence suggesting the Blue Bus Company was responsible. The base-rate evidence again specified that 85% of the buses in the area were owned by the Blue company, and that 15% were owned by the Green company. The witness evidence now came from a single witness who said the accident was caused by a blue bus. This witness was described as accurate at differentiating the buses 85% of the time, and inaccurate 15% of the time. After reading about each kind of evidence, participants used a 7-point Likert scale to rate how likely the judge would be to convict the Blue Bus Company and force them to pay damages ("How likely [is JudgeComp / is Judge Brown / are you] to convict the Blue Bus Company and force them to pay damages"). This scale ranged from "Extremely Unlikely" (1) to "Extremely Likely" (7).

Experiment 3 was identical to Experiment 2, except for two changes: First, we removed the condition where participants served as the judge. Second, rather than contrasting

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press



(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, an autonomous computer system, is responsible for determining the verdict. *Will JudgeComp convict the Blue Bus Company for killing the dog?*



(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, an autonomous computer system, is responsible for determining the verdict.

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?*

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 1: Sample stimuli from Experiment 1 (top) and Experiment 2 (bottom). The top panel shows stimuli from the base-rate condition in Experiment 1. Presentation order of the final two slides was randomized. The bottom panel shows stimuli for the computer judge condition from Experiment 2. Information and questions about each kind of evidence appeared on further slides without images, with presentation order again randomized.

percentages of 85% and 15%, the evidence now contrasted 98% and 2%.

Finally, Experiment 4 used the same procedure as Experiment 3, except the evidence contrasted the percentages 80% and 20%, and participants responded to a yes/no question about whether the Blue company would be convicted. Both of these changes were adopted to more closely match Wells' (1992) original methods.

2.2 Results

Figure 2 shows ratings across the four experiments. In each analysis, the predictors were evidence and judge type. As we detail below, the overall pattern of findings was the same in each experiment. In each, we observed significant main effects of both predictors: Participants were more likely to predict a guilty verdict when the case rested on witness testimony than base-rates, and when the verdict was decided by a computer judge rather than a human. The interaction between these factors, though, was always non-significant.

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

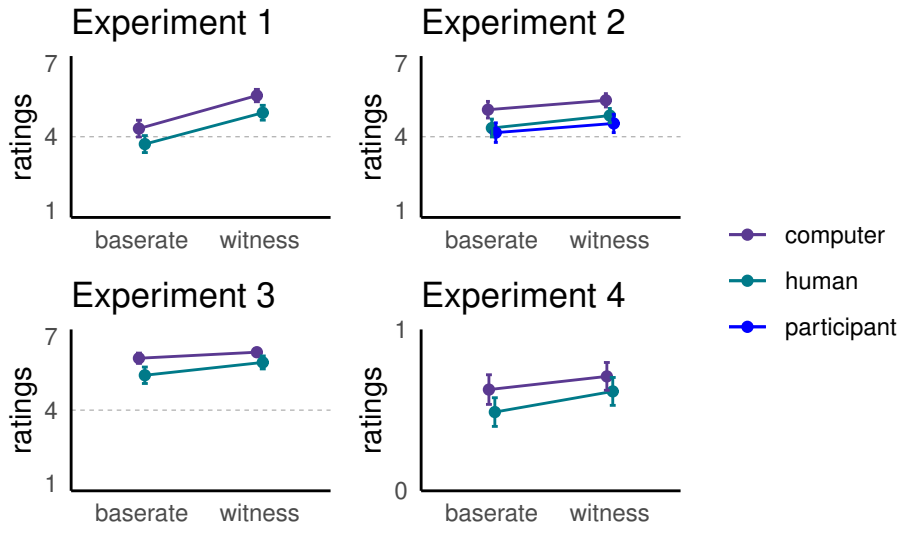


FIGURE 2: Mean conviction ratings in Experiments 1 to 4. Error bars show 95% confidence intervals.

2.2.1 Experiment 1

The ANOVA revealed a main effect of evidence type, $F(1, 232) = 40.41$, $p < .001$, $\eta_p^2 = .15$, a main effect of judge type, $F(1, 232) = 52.85$, $p < .001$, $\eta_p^2 = .19$, but no significant interaction, $F(1, 232) = 0.14$, $p = .710$, $\eta_p^2 < .01$.

2.2.2 Experiment 2

The ANOVA revealed a main effect of evidence type, $F(1, 314) = 31.99$, $p < .001$, $\eta_p^2 = .09$, a main effect of judge type, $F(2, 314) = 9.02$, $p < .001$, $\eta_p^2 = .05$, and no significant interaction, $F(2, 314) = 0.34$, $p = .713$, $\eta_p^2 < .01$. This experiment contrasted judgments for three kinds of judges (manipulated between subjects): A computer, a human, and the participant themselves. Pairwise comparisons (Tukey method) revealed that participants gave greater predictions of a guilty verdict in the condition where the judge was a computer ($M = 5.30$) than in the conditions where the judge another human ($M = 4.61$; $p = .007$), or the participant themselves ($M = 4.36$; $p = .002$). There was no significant difference in ratings between the latter two condition, $p = .497$.

2.2.3 Experiment 3

The ANOVA revealed a main effect of evidence type, $F(1, 234) = 26.60$, $p < .001$, $\eta_p^2 = .10$, and a main effect of judge type, $F(1, 234) = 11.31$, $p < .001$, $\eta_p^2 = .05$, $p = .023$, but no significant interaction, $F(1, 234) = 3.49$, $p = .063$, $\eta_p^2 = .01$.

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

2.2.4 Experiment 4

The generalized estimating equations model revealed a main effect of evidence type, F (1) = 10.97, p = .001, a main effect of judge type, F (1) = 4.59, p = .032, but no significant interaction, F (1) = 0.53, p = .466.

2.3 Discussion

In sum, participants thought that a guilty verdict would be more likely when evidence came from eye-witness testimony than base-rates, and when the case was decided by a computer rather than a human judge. Contrary to our expectations, though, these factors did not interact — participants did *not* expect the computer judge to give relatively more weight than the human judge to base-rate information (or relatively less weight to the individuating eyewitness testimony).

Although the findings did not support our original hypothesis, they nonetheless suggest differences in people's beliefs about how computers and humans reach decisions. One explanation for our findings is that participants expected the computer judge would commit to punitive decisions based on less evidence than a human would require. Such hair-trigger commitment could have worrisome consequences, if it happened in real life. In legal decisions, it might be expected to lead to rash decisions about who should be arrested and convicted, and in military decisions it might lead to rash decisions about who should be considered an enemy and targeted for fire. Another possibility, however, 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 to impose punishments, in each case based on less information than a human judge would require. We tested these accounts in our next experiment.

3 Experiment 5

Participants read about a taxi that either harmed or helped a woman: it either ran a woman over and badly injured her, or it rushed an already-injured woman to a hospital. Some evidence connected the Blue Taxi Company with the incident. As in the previous experiments, this either came from witness testimony or from base-rates. In the harm version, participants predicted whether a computer judge and a human judge would each punish the Blue company. In the help version, participants predicted whether these judges would reward the taxi company.

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

3.1 Method

3.1.1 Participants

We tested 707 participants (M_{age} = 41 years, SD = 13.36; range = 19–89; 39% female); an additional 148 participants were excluded. We recruited a larger sample for this experiment because it used a 2X2X2 design, whereas the previous experiments featured 2X2 designs.

3.1.2 Procedure

This experiment used a slightly different method than those previous. For example, rather than conveying vignettes with images across a series of slides, the vignette was now a written story with no images. Also, we changed the story so that it was about an incident involving a taxi, rather than a bus; see Figure 3 (left). These changes were adopted mostly so that we could easily examine judgments about comparable harmful and helpful outcomes.

One night a taxi [ran over a woman, badly injuring her / rushed an injured woman to a hospital].

Only the Blue Taxi Company and the Green Taxi Company operate in the area.

The woman recovered but could not identify the color of the taxi that [hit / helped] her.

However, [85% of taxis in the area belong to the Blue Company, and only 15% belong to the Green Company. / a witness said the taxi was from the Blue Company. At night, he accurately identifies taxi colors 85% of the time. He is inaccurate 15% of the time.]

Suppose [**JudgeComp, an autonomous computer system,** / **Judge Brown, a 56-year-old from Chicago**] is responsible for determining if the Blue Taxi company should be [punished for hitting / rewarded for helping] the woman.

How likely is [**JudgeComp / Judge Brown**] to [punish / reward] the Blue Taxi Company?

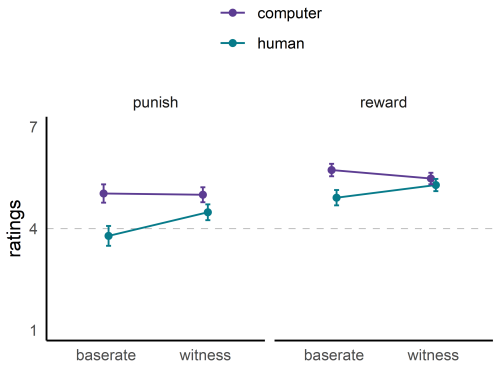


FIGURE 3: Script and ratings in Experiment 5. In the script, square brackets show text manipulated across conditions. Each participant was asked test questions about both judges (order randomized across participants). In the plot, error bars show 95% confidence intervals.

Event type (harmful, helpful) and evidence-type (base-rates, witness) were both manipulated between subjects. In the harmful condition, the taxi ran over a woman, injuring her; in the helpful condition, the woman was already injured and the taxi rushed her to the hospital. Judge type (computer, human) was manipulated within subjects. The test question for each judge was presented on a separate screen and presentation order was random. Participants indicated the likelihood that each judge would punish or reward the Blue company using a 7-point Likert scale ranging from "Very Unlikely" (1) to "Very Likely" (7).

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

3.2 Results

Figure 3 (right) shows participants' ratings. An ANOVA with the predictors evidence type, judge type, and event type revealed main effects of evidence type, $F(1, 703) = 4.02$, $p = .045$, $\eta_p^2 < .01$, judge type, $F(1, 703) = 152.88$, $p < .001$, $\eta_p^2 = .18$, and event type, $F(1, 703) = 62.83$, $p < .001$, $\eta_p^2 = .08$. Participants thought the computer was overall more likely than the human judge to decide based on the evidence. They also thought judges were overall more likely to reward the taxi company (woman was helped) than to punish it (woman was harmed).

The analysis also revealed significant 2-way interactions between event type and judge type, $F(1, 703) = 11.44$, $p < .001$, $\eta_p^2 = .02$, and between evidence type and judge type, $F(1, 703) = 36.49$, $p < .001$, $\eta_p^2 = .05$. The interaction between event type and evidence type was non-significant, $F(1, 703) = 1.88$, $p = .171$, $\eta_p^2 < .01$, as was the 3-way interaction, $F(1, 703) = 0.24$, $p = .622$, $\eta_p^2 < .01$.

The event type X judge type interaction resulted because although participants thought the computer would be overall more likely than the human to decide based on the evidence, this effect was greater when punishing the company, $M_{difference} = 0.88$, $SE = 0.08$, $p < .001$, than when rewarding it, $M_{difference} = 0.50$, $SE = 0.08$, $p < .001$. As we revisit below, this finding supports *both* explanations for why participants in Experiments 1 to 4 predicted the computer would be more likely than the human judge to convict.

The evidence type X judge type interaction resulted because participants thought the human judge would be more like to decide based on witness testimony than base-rates, $M_{difference} = 0.53$, $SE = 0.12$, $p < .001$, whereas ratings for the computer judge did not significantly vary based on evidence type, $M_{difference} = -0.14$, $SE = 0.11$, $p > .183$. These findings contrast the findings from Experiments 1 to 4 and are consistent with our original hypothesis.

3.3 Discussion

Participants thought that computers would be overall more likely than humans to base decisions on the uncertain evidence available. They also thought this difference would be especially apparent for decisions to punish. As such, the findings separately support *both* explanations we had offered for why participants in Experiments 1 to 4 thought the computer judge would be more likely than the human to punish. Also, participants now expected humans to give relatively more weight than computers to individuating evidence, compared with evidence from base-rates. This was the hypothesis that originally motivated the present research, but it was not supported in Experiments 1 to 4.[3]

[3]These findings came as a surprise to us! By the time we conducted this experiment, we had given up on our original hypothesis, and were mainly trying to better understand why participants predicted the computer judge would be more likely than the human one to reach a guilty verdict.

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

Why did participants in this experiment, but not in any of the previous four, expect the computer judge to be insensitive to the difference between individuating and base-rate evidence? This could come down to methodological differences between this experiment and the first four. For example, participants in the earlier experiments saw cartoon-like images of the judges, whereas in this experiment, participants read text only. The cartoon-like images may have prompted participants to treat the scenarios as purely fictional stories, a context where human-like computers might seem reasonable. So, participants who saw images may have anthropomorphized the computer, viewing it somewhat more like a human agent than a real computer. Also, participants in the earlier experiments read about a less serious accident than did participants in the "punish" condition of this experiment (i.e., a dog being killed versus a woman being injured). Perhaps this difference also somehow impacts participants' expectations about computer judges. We conducted a final experiment to examine whether either of these methodological differences might impact participants judgments.

4 Experiment 6

Participants read about a taxi that either injured a woman or killed her dog, and they either saw cartoon-like images of the judges or read text only. As in the earlier experiments, the Blue Taxi Company was suspected because of witness testimony or base-rates.

4.1 Method

4.1.1 Participants

We tested 935 participants (M_{age} = 40 years, SD = 12.65; range = 19–80; 48% female); an additional 65 participants were excluded.

4.1.2 Procedure

Participants read a vignette about a taxi accident in a 2X2X2X2 design. Three factors were manipulated between subjects: whether the accident was more or less serious (i.e., woman injured versus dog killed); whether participants saw cartoon-like images of the judges or read text only; and whether the Blue Bus Company was suspected because of witness testimony or base-rates. Judge type (computer, human) was manipulated within subjects. Figure 4 (left) shows a screenshot of the vignette and text question about the computer judge in the condition where participants saw cartoon-like images. Participants again indicated the likelihood that each judge would punish the Blue company using a 7-point Likert scale ranging from "Very Unlikely" (1) to "Very Likely" (7).

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

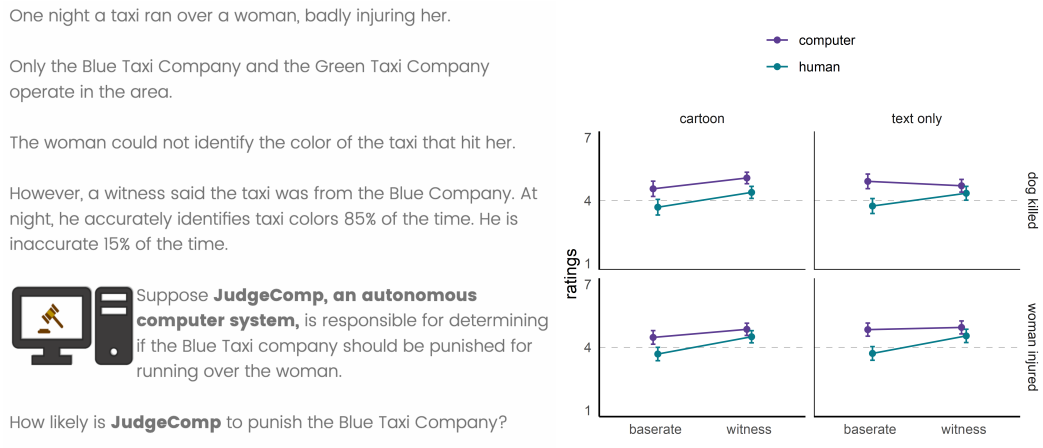


FIGURE 4: Sample stimuli and ratings from Experiment 6. The left panel shows a screenshot from the condition where participants saw cartoon-like images of each judge. The text-only version was identical except the image did not appear, so the text was not indented. In the plot, error bars show 95% confidence intervals.

4.2 Results

Figure 4 (right) shows participants' ratings. As preregistered, we examined the results using two separate 2X2X2 analyses. Both analyses included evidence type and judge type as predictors, so they only differed in the remaining predictor. In one analysis, this was format (cartoon, text only); in the other analysis, it was seriousness (dog killed, woman injured).

4.2.1 ANOVA with format

This analysis revealed main effects of evidence type, $F(1, 931) = 21.62$, $p < .001$, $\eta_p^2 = .02$, and judge type, $F(1, 931) = 196.82$, $p < .001$, $\eta_p^2 = .17$, but no main effect of format, $F(1, 931) = 0.42$, $p = .519$, $\eta_p^2 < .01$. Participants again thought punishment was more likely when evidence was witness testimony than base-rates, and when the judge was a computer rather than a human. The analysis also revealed a significant 2-way interactions between evidence type and judge type, $F(1, 931) = 26.72$, $p < .001$, $\eta_p^2 = .03$, but the other 2-way interactions were non-significant: evidence type by format, $F(1, 931) = 1.83$, $p = .177$, $\eta_p^2 < .01$, and cartoon type by judge type, $F(1, 931) = 0.70$, $p = .402$, $\eta_p^2 < .01$. The 3-way interaction was also significant, $F(1, 931) = 5.38$, $p = .021$, $\eta_p^2 < .01$.

To better understand the 3-way interaction, we examined whether participants were sensitive to the distinction between witness testimony and base-rates for each of the four combinations of judge and format. Participants thought the human judge would be more likely to punish based on witness testimony than base-rates, both when they saw cartoons, $M_{difference} = 0.76$, $SE = 0.16$, $p < .001$, and when they read text only, $M_{difference} = 0.72$, $SE = 0.17$, $p < .001$. In contrast, whereas participants who saw cartoons also expected the computer judge to be sensitive to the type of evidence, $M_{difference} = 0.47$, $SE = 0.16$, $p =$

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

.003, participants who saw only text did not expect this, $M_{\text{difference}} = 0.05$, $SE = 0.16$, $p = 0.760$. Together, this suggests variations in findings across the earlier experiments resulted because participants in Experiments 1 to 4 saw cartoon-like images of the computer judge, whereas participants in Experiment 5 did not.[4]

4.2.2 ANOVA with seriousness

This analysis again revealed main effects of evidence type, $F(1, 931) = 21.28$, $p < .001$, $\eta_p^2 = .02$, and judge type, $F(1, 931) = 195.71$, $p < .001$, $\eta_p^2 = .17$, and no main effect of seriousness, $F(1, 931) = 0.02$, $p = .888$, $\eta_p^2 < .01$. All interactions involving seriousness were non-significant: evidence type by seriousness $F(1, 931) = 0.41$, $p = .524$, $\eta_p^2 < .01$; judge type by seriousness, $F(1, 931) = 1.24$, $p = .267$, $\eta_p^2 < .01$; 3-way interaction, $F(1, 931) = 0.16$, $p = .691$, $\eta_p^2 < .01$. Hence, the only significant interaction was the 2-way interaction between evidence type and judge type, $F(1, 931) = 26.93$, $p < .001$, $\eta_p^2 = .03$. It resulted because participants thought the human judge would be more like to decide based on witness testimony than base-rates, $M_{\text{difference}} = 0.74$, $SE = 0.12$, $p < .001$, whereas ratings for the computer judge did not significantly vary based on evidence type, $M_{\text{difference}} = 0.20$, $SE = 0.11$, $p = .073$.

4.3 Discussion

Participants again expected humans to be more sensitive than computers to the distinction between witness testimony and base-rates. But this sensitivity depended on the way vignettes were presented, and specifically on whether participants read text only, or instead saw a cartoon-like image of the computer judge. Participants' judgments were not affected, though, by the seriousness of the accident in the vignette. Also, as in all the previous experiments, participants again thought a computer judge would be overall more likely than a human one to punish.

5 General Discussion

We examined whether people expect machines to differ from humans in how they make decisions. In our experiments, participants predicted whether computer and human judges would find a transit company guilty for causing an accident. Our original hypothesis was that people would expect human and computer judges to differ in their treatment of individuating information and base-rates. Besides finding support for this hypothesis (though with some

[4]This conclusion was further suggested by separate 2(format: cartoon, text only) X 2(evidence type: witness, base-rates) analyses conducted for each type of judge. With the human judge, there was a main effect of evidence type, $p < .001$, but no main effect of format and no interaction, $ps \geq .845$. But with the computer judge, the interaction was significant, $p = .023$, and the main effects were not, $ps \geq .066$. This again suggests that the format impacted people's expectations about the computer judge.

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

caveats), our studies also uncovered other ways people expect humans and computers to differ when making decisions.

Overall, our findings support the idea that people expect humans and computers to differ in how they prioritize different kinds of information. Specifically, in the final two experiments, participants anticipated that a human judge would be more swayed by individuating information than base-rates, while expecting a computer judge to treat these two kinds of evidence equivalently. This finding fits with the possibility that people often expect computers to make numerical and statistical decisions without understanding what the numbers and statistics mean, or where they come from.

This said, participants' responses only fit this belief when the vignettes did not include cartoon-like images of the computer judge. When images were included, participants predicted that the computer and human judges would be similarly sensitive to the distinction between individuating evidence and base-rates. We suspect that the cartoon-like image encouraged participants to treat the scenario as purely fictional stories. In stories, human-like computers might seem reasonable and so this might diminish expectations about how decisions by humans and computers will differ. If this account has merit, we might likewise expect predictions of computer and machine decision-making to vary depending on how much the machines resemble humans, as this strongly impacts whether machines (and other non-human agents) are likely to be anthropomorphized and treated as having mental states (for reviews see Marchetti et al., 2018 and Waytz et al., 2013).

Participants also expected computer and human decision making to differ in other ways. One other difference is that participants predicted computers would be more likely than humans to make consequential decisions based on probabilistic evidence. This difference between expectations about computers and humans was apparent in all six experiments and was robust across many manipulations, both within and across experiments, and it held up regardless of whether participants saw images of the judges or read text only. We saw this difference for decisions about both punishment and reward; for individuating evidence and base-rate evidence; for evidence with more moderate odds (80%) and very high odds (98%); for responses to both Likert scales and yes/no questions.

Another related difference is that participants predicted computers would be more likely than humans to punish. This finding was observed in the fifth experiment, which compared predictions about reward and punishment. Whereas participants predicted the human judge would be much less likely to punish than reward, they predicted the computer would be just a bit less likely to punish than reward. It is worth acknowledging, though, that this difference may have resulted because the harmful act in our vignette was viewed as more extreme than the helpful one — the negative valence of being guilty of a hit-and-run is probably greater than the positive valence of voluntarily driving an injured woman to a hospital.

Both of these differences between expectations about computers and humans could result from the same underlying belief. Specifically, people might believe that computers lack the metacognitive capacity to feel uncertain — a belief that may be accurate for now (Fleming,

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

2021a, pp. 193-203; Fleming, 2021b). Further evidence that people hold this belief comes from findings that people explicitly deny that computers and robots are conscious and self-aware (Weisman et al., 2017). Feeling doubt and uncertainty may be important for refraining from committing to decisions, especially when the cost of a potential error is high — as is true with wrongfully punishing someone (see Zamir & Ritov, 2012). Hence, participants may have predicted the computer judge would decide based on the evidence because they thought it could not be deterred by doubt.

Together, our findings are informative about people's beliefs about how machines make decisions, and how this differs from human decision making. Hence, the findings are informative about people's theory of machine (Logg et al., 2019), though they could also be characterized as informing us about people's understanding of how machine minds differ from human ones. The present experiments also initiate a new connection between 'theory of mind' and judgment and decision making. Research on theory of mind examines the abilities to infer how humans and other agents think and feel, and to predict how they will act in light of these mental states. But, research in theory of mind has not often capitalized on distinctions discovered by judgment and decision making researchers, such as contrasts between individuating and base-rate information.

Our findings may also relate to algorithm aversion — people's misgivings about entrusting computers and other machines with certain decisions (for an overview see Burton et al., 2020). People are often averse to entrusting machines with moral decisions (Bigman & Gray, 2018), decisions viewed as subjective and as requiring intuiton (Castelo et al., 2019), and decisions where it seems important to consider factors that have elements which make them unique and distinctive (Longoni et al., 2019). People also avoid relying on machines after seeing them make mistakes, even when the decision to instead rely on a human prevents people from maximizing earnings (Dietvorst et al., 2015).

Although we did not examine algorithm aversion, our findings raise the possibility that misgivings about machine decision making may sometimes result in concerns that machines will decide based on weaker evidence than most humans would feel is necessary. This concern might not strongly affect people's trust in allowing machines to make decisions that confer benefits or assign rewards. But the concern could make people weary about entrusting algorithms with decisions that have the potential to cause harm. If so, we might expect to see an asymmetry in which moral decisions are most affected by algorithm aversion.

References

Arkes, H. R., Shoots-Reinhard, B., & Mayes, R. S. (2012). Disjunction between probability and verdict in juror decision making. *Journal of Behavioral Decision Making*, *25*, 276–294. https://doi.org/10.1002/bdm.734.

https://doi.org/10.1017/S1930297500003600 Published online by Cambridge University Press

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44,* 211–233. https://doi.org/10.1016/0001-6918(80)90046-3.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181,* 21–34. https://doi.org/10.1016/j.cognition.2018.08.003.

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. https://doi.org/10.1002/bdm.2155.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788.

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.

Flanagan, T., Rottmann, J., & Howard, L. H. (2021). Constrained choice: Children's and adults' attribution of choice to a humanoid robot. *Cognitive Science*, *45*, e13043. https://doi.org/10.1111/cogs.13043.

Fleming, S. M. (2021a). *Know thyself: The science of self awareness*. Basic Books.

Fleming, S. M. (2021b, April 6). *What separates humans from AI? It's doubt.* Financial Times Magazine. https://www.ft.com/content/1ff66eb9-166f-4082-958f-debe84e92e9e.

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

Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, inhuman, and superhuman: Contrasting humans with nonhumans in three cultures. *Social Cognition*, *26*(2), 248–258. https://doi.org/10.1521/soco.2008.26.2.248.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251. https://doi.org/10.1037/h0034747.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*, 143–157. https://doi.org/10.1016/0010-0277(82)90023-3.

Kim, T. W., & Duhachek, A. (2020). Artificial intelligence and persuasion: a construal-level account. *Psychological Science*, *31*(4), 363–380. https://doi.org/10.1177/0956797620904985.

Koehler, J. J. (2001). When are people persuaded by DNA match statistics? *Law and Human Behavior*, *25*(5), 493–513. https://doi.org/10.1023/A:1012892815916.

Laakasuo, M., Herzon, V., Perander, S., Drosinou, M., Sundvall, J., Palomäki, J., & Visala, A. (2021). Socio-cognitive biases in folk AI ethics and risk discourse. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00060-5.

Lagnado, D. A., & Sloman, S. A. (2004). Inside and outside probability judgment. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 157–176). Malden, MA: Blackwell.

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. https://doi.org/10.1016/j.obhdp.2018.12.005.

Logg, J. (2021). The psychology of big data: Developing a "theory of machine" to examine perceptions of algorithms. To appear in Matz, S.(Ed.), *American Psychological Association Handbook of Psychology of Technology*.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4)*,* 629–650. https://doi.org/10.1093/jcr/ucz013.

Marchetti, A., Manzi, F., Itakura, S., & Massaro, D. (2018). Theory of mind and humanoid robots from a lifespan perspective. *Zeitschrift für Psychologie, 226*(2), 98–109. https://doi.org/10.1027/2151-2604/a000326.

Niedermeier, K. E., Kerr, N. L., & Messé, L. A. (1999). Jurors' use of naked statistical evidence: Exploring bases and implications of the Wells effect. *Journal of Personality and Social Psychology, 76*(4), 533-542. doi:10.1037/0022-3514.76.4.533.

Searle, John. R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences*, *3(3)*, 417–457. https://doi.org/10.1017/S0140525X00005756.

Turri, J., Friedman, O., & Keefner, A. (2017). Knowledge central: a central role for knowledge attributions in social evaluations. *Quarterly Journal of Experimental Psychology*, *70*(3), 504–515. https://doi.org/10.1080/17470218.2015.1136339.

Tversky, A., & Kahneman, D. (1977). Causal thinking in judgment under uncertainty. In R.E. Butts & J. Hintikka (Eds.), *Basic problems in methodology and linguistics* (pp. 167–190). Springer.

Waytz, A., Klein, N., & Epley, N. (2013). Imagining other minds: Anthropomorphism is hair-triggered but not hare-brained. In M. Taylor (Ed.), *The Oxford handbook of the development of imagination* (pp. 272–287). Oxford University Press.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, *114*(43), 11374–11379.

Wells, G. L. (1992). Naked statistical evidence of liability: Is subjective probability enough? *Journal of Personality and Social Psychology, 62,* 739–752. https://doi.org/10.1037/0022-3514.62.5.739.

Zamir, E., & Ritov, I. (2012). Loss aversion, omission bias, and the burden of proof in civil litigation. *The Journal of Legal Studies*, *41*(1), 165–207. https://doi.org/10.1086/664911.