When Do People Want AI to Make Decisions?

Max F. Kramer
University of Arizona

Jana Schaich Borg, Vincent Conitzer, Walter Sinnott-Armstrong
Duke University

Abstract

AI systems are now or will soon be sophisticated enough to make consequential decisions. Although this technology has flourished, we also need public appraisals of AI systems playing these more important roles. This article reports surveys of preferences for and against AI systems making decisions in various domains as well as experiments that intervene on these preferences. We find that these preferences are contingent on subjects' previous exposure to computer systems making these kinds of decisions, and some interventions designed to mimic previous exposure successfully encourage subjects to be more hospitable to computer systems making these weighty decisions.

Introduction

AI is playing an ever more significant role in transportation, industry, war, finance, healthcare, and other domains. One may ask whether machines should be allowed to make decisions in these domains for us. This question becomes most pressing when decisions have significant ethical dimensions. It is one matter to produce AI that plays chess and another to produce AI that carries out drone strikes on humans. Further, it is one matter to produce AI that carries out a human-specified drone strike and another to give AI control over where the bombs fall. In this study, we ask people whether they prefer humans or computers to make decisions with important consequences in various scenarios. Then we intervene on aspects affecting people's preferences for computers making moral decisions and glean insights from the success and failure of the various interventions. We find that preferences for computer decision-makers are contingent on prior exposure to computers performing those kinds of tasks and find that some interventions that mimic the effect of prior exposure of this kind are effective at shifting preferences.

Background

The public image of AI has been shaped by programs that beat humans at Jeopardy, chess, and Go, but these successes are not the most consequential for our lives. Consider this example from the AAAI Conference on Innovative Applications of Artificial Intelligence: a program that can diagnose common cognitive disorders (Davis et al. 2015). A computerized psychiatrist would constitute a considerable change in many people's daily lives as well as their conception of medical practice writ large.

The question of how AI should act in high-stakes decisions is one that has been receiving increasing attention. One common strain of such work involves applying the structure of thought experiments such as the trolley problem (originating in Foot 1967) to AI systems such as autonomous vehicles. Imagine an autonomous vehicle faced with a dilemma: it can either drive off the road into a barrier, killing the passengers in the car, or drive into some number of persons on the road, killing them (e.g., Bonnefon, Shariff, and Rahwan 2016). This problem is difficult enough when a person is driving the car, but the automation of the vehicle requires its program to account for these kinds of cases (whether explicitly or implicitly) before the cars are put on the road. Along these lines, some researchers have investigated how robots and AI systems should be programmed to make moral decisions (e.g., Bonnefon, Shariff, and Rahwan 2016; Wallach and Allen 2009; Freedman et al. 2018; Noothigattu et al. 2018), granting them the ability to make those kinds of moral decisions.

This question becomes crucial as the scope of AI expands within our society. We need to decide not just whether we *can* build AI to make moral decisions, but also whether 'moral AI' systems *should be allowed* to make decisions in a given context at all. Take, for instance, the paper by Davis and colleagues (2015) referenced above. Even if AI can be implemented in a role normally performed by a clinician, should it? Is it acceptable to implement AI in settings where most stakeholders do not trust AI and instead want humans to make decisions? In the studies that follow, we provide insight into when and why people prefer AI systems to play consequential roles in their lives in situations such as these.

Studies 1a-c

Methods and Materials

To set out, we wanted to first see whether there were any individual factors that were related to whether people prefer humans or computers to make decisions, across a variety of contexts. One might think that certain people, owing to personality traits they have or demographic traits that may im-

pact their specific relationship to technology, would be more or less likely to see encroachment of AI into their daily lives as a positive or negative phenomenon.

Subjects were first told whether the computer agents they were to consider had an option for human override or not and were then presented with 18 scenarios that described situations where AI is already being applied or where AI is likely to be applied in the future. For illustration, here is the description of a scenario related to sentencing defendants who are found guilty. "After someone has been found guilty of a crime in the American legal system, a hearing is held to determine what punishment sentence (or time in prison) the guilty party will receive. Information about the guilty party and the circumstances of the crime is usually used to decide what the criminals sentence should be. A new country is writing its constitution after a revolution, and must establish a method for how its legal punishment sentences will be determined after guilty sentences are delivered by a judge or jury." Subjects were then asked to complete sentences of the form "Decisions to [type of decision in question] should be made by..." and "The best decisions about [type of decision in question] are most likely to be made by..." with a selection from a seven-point scale with "definitely computers" at the low end and "definitely humans" at the high end.

Subjects were recruited through Amazon Mechanical Turk (MTurk). Study 1a collected data from 98 participants (45 female), Study 1b collected data from another 98 participants (50 female), and Study 1c collected data from 100 subjects (50 female). In Studies 1b and 1c, in addition to the questions described above, participants were also asked to report the degree to which they had prior experience with computers and humans making the type of decision in question ("How much have you heard about, or had experience with, [computers that/humans who] can perform the type of task described in this scenario"). In all studies, subjects were asked to fill out a variety of demographic and psychometric scales (Study 1a: Social and Economic Conservatism Scale (SECS; Everett 2013), a measure of political preference, Moral Identity Scale (MIS; Aquino and Reed II 2002), a measure of how important certain moral traits are to one's self-image, Risk Propensity Scale (RPS; Meertens and Lion 2008), a measure of risk-seeking/aversion, and the Disgust Scale Revised (DS-R; Haidt, McCauley, and Rozin 1994), a measure of trait disgust; Study 1b: Moral Identity Scale, Risk Propensity Scale, Moral Foundations Questionnaire (MFQ; Graham et al. 2011), a measure of which moral concepts a participant typically employs and cares about, Adapted Empathy Questionnaire (AEQ; Beadle et al. 2015), a measure of trait empathy, Self-Report Psychopathy Scale (SRPS; Hare, Harpur, and Hemphill 1989), a measure of psychopathic traits; Study 1c: Moral Foundations Questionnaire, Adapted Empathy Questionnaire, Arnett Inventory of Sensation-Seeking (AISS; Arnett 1994), a measure of predilection to seek out high-arousal situations, and Abbreviated Impulsiveness Scale (ABIS; Coutlee et al. 2014), a measure of trait impulsiveness).

Results

To analyze the data, we first created a summary variable of average preference across all scenarios as well as overall scores for each psychometric survey. We ran correlation analyses between psychometric scores and continuous demographic variables and the average preference measure. We also ran one-sample t-tests to determine on which scenarios participants expressed preferences significantly different from the midpoint of the scale, which represented a 'not sure' answer, and one-way ANOVAs to determine whether overridability condition generated a main effect of preference.

In Study 1a, we found that ratings of how good computers and humans were at performing complicated tasks (not specific to any of the tasks in the study) correlated positively with preferences for computers and humans, respectively (p=.008 for computers and p=.006 for humans). Also, participants' scores on the MIS correlated positively with participants choosing computers over humans to complete tasks (p=.007). No other demographic or psychometric items (SECS, RPS, and DS-R) correlated significantly with the summary dependent variables (the mean of preferences on each scenario). Surprisingly, the overridability conditions did not make a significant difference in overall preference (p>.1). In other words, whether computers' decisions could be overridden by humans did not have a significant effect on whether participants chose computer decision-makers over human decision-makers or vice versa.

Study 1a was designed to examine participants' general preferences for human or computer decision-makers across a wide variety of contexts, rather than to uncover details about how participants respond to any specific scenario. However, we noticed that the scenarios where participants were more likely to prefer computer decision-makers over human decision-makers tended to be ones where subjects may have previously heard or read about computers making those kinds of decisions—for example, choosing the following distance of a vehicle and choosing which advertisements will be shown to consumers. Thus, in Studies 1b and 1c, we tested explicitly whether previous experience with computers making decisions in specific contexts correlated with the participant's preferences for computer decision-makers in those contexts.

Study 1b replicated the principal effects of Study 1a. Despite adding reminders before each preference request, the overridability condition remained an insignificant factor (p>.1). Again, perceived task ability (the response to the query of how well the participant thought computers and humans, respectively, performed complicated tasks) correlated highly with overall preferences (p=.001 for computer and p<.001 for human). However, unlike in Study 1a, moral identity did not correlate with decision-maker preferences (p>.3). In addition, the more familiar participants were with computers making decisions in specific scenarios, the more likely they were to prefer computer decision-makers over human decision-makers in those scenarios (p<.001 for each). When each scenario was examined individually, (i.e., comparing the familiarity on scenario X with the preference on scenario X), familiarity with humans was positively cor-

related with preference for humans in 11 of the 18 scenarios and familiarity with computers was positively correlated with preference for computers in 16 of the 18 scenarios. Furthermore, aggregate previous familiarity with computers (that is, averaged across all scenarios rather than scenario-matched) was significantly negatively correlated with overall preference (p=.03; because preferences for computers constituted the bottom half of the scale, 'higher' overall preferences correspond to greater preferences for humans, while 'lower' overall preferences correspond to greater preferences for computers, and therefore negative correlations with the overall preference indicate a positive correlation with preference for computers). Finally, participant scores on the overall purity scale, both purity subscales, and one authority subscale of the MFQ were significantly positively correlated with overall preference for computers (all p<.04). These scales and subscales reflected the participant's tendency to rely on considerations of moral purity and appeals to authority in their moral judgment.

Study 1c replicated the primary findings of the previous two studies. Again, overridability produced no significant differences on any decision measures. Again, the macro-level within-scenario correlation between both familiarity measures and preferences was highly significant (p<.001), though only aggregate human familiarity, and not aggregate computer familiarity, was significantly correlated with overall preference (p=.047). Within scenarios, previous familiarity with humans correlated significantly on 10 of 18 scenarios and previous familiarity with computers correlated significantly on 7 of 18 scenarios (all p<.05). The effects from Study 1b relating some of the moral foundations to overall preference were not replicated. Across Studies 1a-1c, then, we see that prior exposure to computers is the only item (aside from a general appraisal of how well computers and humans can perform complicated tasks) that is consistently correlated with a preference for computers in weighty decision-making contexts.

Studies 2a and 2b

Methods and Materials

In Study 2a and 2b, we examined whether the discovered relationship between exposure and preference is causal by manipulating participants' preferences for computer decision-makers. We did this by exposing them to information about computers completing related tasks. In this set of studies, we focused primarily on one scenario about an agent dictating the process by which patients in a kidney exchange would be selected for inclusion in a given round of the exchange. In a kidney exchange, patients in need of a kidney transplant who have a willing but less than ideal—e.g., medically incompatible—donor attempt to exchange their donors. We focused on this scenario because average preferences on the kidney exchange scenario hovered around the midpoint of the scale in Studies 1a-1c, reflecting uncertainty about which type of agent should be responsible for the decision; the prior exposure of participants to this situation was very low for both computers and humans; and it is in fact a domain where AI is already used (see, for instance, Dickerson and Sandholm 2015), allowing us to truthfully inform participants of this. All this together made this scenario an ideal target for an exposure intervention. We created four interventions in the form of articles about the kidney exchange process adapted from literature produced by the National Kidney Registry. The articles first gave a general description of the kidney exchange process, including a diagram of a kidney exchange between three donor-patient pairs. They then mentioned that either computers or humans could take the role of directing a kidney exchange. Finally, some of the articles provided additional information that was intended to replicate different components of prior exposure to computers making decisions in a kidney exchange environment.

All variants of the intervention had the same core article, but they differed in sentences put at the end of the article. The variants of the intervention were created in a 2 x 2 on-off format, with two variables corresponding to sentences that were either included or excluded from the end. The "success" sentence read, "It is believed that computers can coordinate exchanges with higher success rates than humans," and the "status quo" sentence read, "Computers are currently being used to make this decision in some exchanges." Each of these variables was meant to capture a different component or interpretation of 'prior exposure': simple knowledge of existence ("status quo") or evaluative knowledge of functioning ("success").

Again, participants (88 in Study 2a, 36 female; 89 in Study 2b, 41 female) were recruited from and participated in the study using MTurk. In a within-subjects design with respect to preference with and without intervention, participants were given six of the scenarios from Studies 1a-c—the kidney exchange scenario, one scenario each that in the experiments was strongly human-preferred, weakly human-preferred, weakly computer-preferred, and strongly computer-preferred, and another scenario in a medical context—and asked for their preferences between human and computer decision-makers as in Studies 1a-1c. Then, each participant was randomly given one of the four interventions, 'Neither' (N), 'Status Quo Only' (SQO), 'Success Only' (SO), or 'Both' (B). After reading the intervention, the participant responded to a question in order to confirm their comprehension of the content. Finally, they were asked again for their preferences on the same six scenarios that they had previously seen. Differences between their initial and post-intervention preferences were treated as dependent variables, i.e., the effect of the various interventions. Finally, the participant completed a demographic survey. We hypothesized that the interventions would create graded shifts in preference for computer agents on the kidney exchange scenario and the other medical scenario, such that the N intervention would create a small shift in preference, SQO and SO a larger shift, and B the largest shift. We also hypothesized no effect of the interventions on the non-medical scenarios.

Results

We used a one-sample t-test to determine the significance of preference changes, both within and across intervention conditions. We then conducted an ANOVA with planned com-

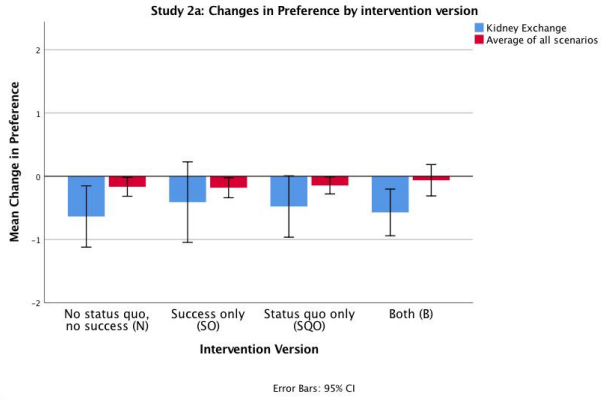

Figure 1: Changes in preference separated by intervention version in Study 2a. Note that in all figures, negative values indicate a shift toward computer preference and positive values indicate a shift toward human preference.

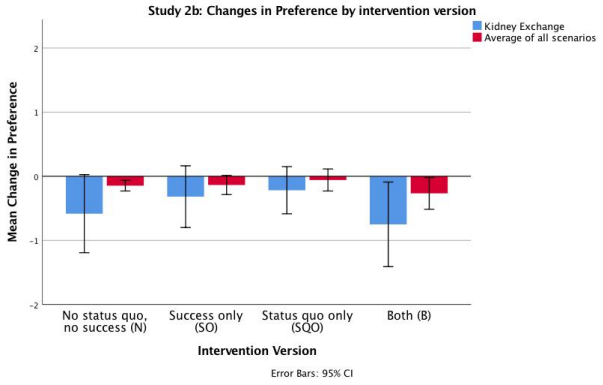

Figure 2: Changes in preference separated by intervention version in Study 2b.

parisons to investigate an effect of intervention condition on preference change.

Overall, collapsing across all four versions of the kidney exchange article, the intervention was successful in shifting preference toward computer decision-makers in the kidney exchange scenario ($p<.001$) and the other medical scenario about prescribing medicine ($p=.034$) as well a scenario about investigating potential cases of bribery ($p=.031$; other scenarios $p>.07$). On average, preferences on the kidney exchange scenario shifted over 0.5 points on a 5-point scale in the direction of preferring computers, while the other two significant effects represented changes of less than 0.2 points. When we partition the dataset by intervention variant, we find that preferences on the kidney exchange scenario change significantly only in the N ($p=.013$) and B ($p=.004$) conditions. Preferences did not change significantly in the SO ($p=.196$) and SQO ($p=.053$) conditions. Thus, we see that participants only significantly responded to a very strong intervention (B) or an intervention with no analog for prior exposure (N), rather than interventions with an analog for only one sort of exposure (SO and SQO). The only intervention that significantly shifted preferences in the bribery scenario was SO ($p=.039$) and the only intervention that significantly shifted preferences in the prescription scenario was SQO ($p=.032$).

In Study 2b, we attempted to replicate this strange effect in a second independent cohort of participants. Again, collapsing across condition, the average participant shifted just under 0.5 points in their preferences on the kidney exchange scenario ($p=.001$) and around 0.2 points in their preferences on the bribery scenario ($p=.027$; all other scenarios $p>.1$). This time, the only intervention that significantly (at the $p=.05$ level) shifted preferences on the kidney exchange scenario was the B intervention. Figures 1 and 2 provide visualizations of these results. The only intervention driving a change in bribery scenario preference was the N intervention ($p=.043$; all other interventions $p>.09$).

Since Studies 2a and 2b are identical in their structure,

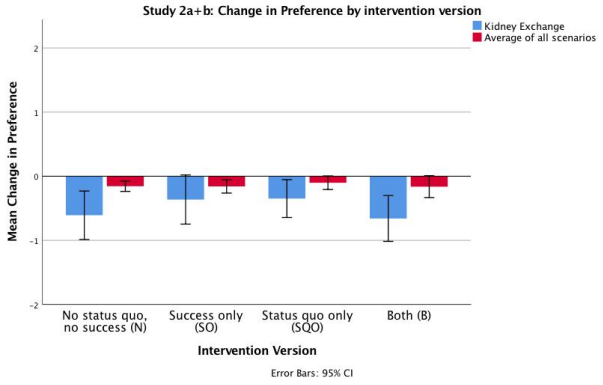

Figure 3: Changes in preference separated by intervention version in the combined data set of Studies 2a and 2b.

we can also combine the two for a more powerful analysis. When we do this, we find mainly the same results. Collapsing across interventions, we see that the average participant shifted their kidney exchange preference by 0.49 points on the 5-point scale ($p<.001$). Once more, ANOVA testing tells us that the N and B interventions produced significant shifts in preference toward computers in the kidney exchange scenario. Unlike the analyses for Study 2a and Study 2b individually, the increased power of Study 2a+b renders the preference change of those who received the SQO intervention significant ($p=.022$) while the effect of the SO intervention nears but does not reach significance ($p=.062$). At a coarse grain, however, the pattern still holds: N and B interventions appear more powerful than SQO and SO (see Figure 3). Across interventions, the bribe ($p=.002$) and prescription ($p=.012$) scenarios saw significant preference change, though the effect is only present in the N and SQO conditions for bribe preference ($p=.017$ and .008, respectively) and only in the B condition for prescription preference ($p=.034$).

Study 3

Methods and Materials

In this study, we wanted to determine whether the manipulation we implemented in Study 2 would hold in a design in which each participant was exposed to the equivalent of each of the kidney exchange interventions. The motivation for this design was to explore why N and B are apparently the most powerful interventions, contrary to intuition, which would predict that the SO and SQO interventions would produce greater preference change than the N intervention. Our hypothesis was that when information about only success or current practice was given, participants would automatically infer the negation of the other type of exposure. The way to test this was to give participants multiple interventions so that they would be aware of different potential kinds of evidence.

Once again, participants give preferences between computer and human decision-makers in six scenarios, one of which deals with the design of a new kidney exchange, and are given a kidney exchange intervention in the form of an article and asked a second time for their preferences on the same scenarios. The variation in Study 3 is that all participants were given the same intervention (N) and then subsequently asked a series of hypothetical questions constituting the other three conditions from Studies 2a and 2b (SQO, SO, and B) for the kidney exchange scenario. Participants were randomly assigned to either the 'explicit' condition (in which their SQO and SO hypotheticals explicitly mentioned the lack of the other type of exposure) or the 'implicit' condition (in which, e.g., the SQO hypothetical makes no mention of the expected success of computers organizing kidney exchanges). To illustrate, the wording of the SQO prompt in the explicit condition is "Imagine that you were told by an authority you trust that computers are already being used to coordinate kidney exchanges. Right now, not enough data is available to indicate whether computers can coordinate kidney exchanges with higher success rates than humans," whereas in the implicit condition, the latter sentence is not included. This study did not include an analog of the N intervention as a dependent variable, because the N intervention does not contain a 'missing piece' to be inferred or made explicit. Following these hypotheticals, which were only given in reference to the kidney exchange scenario, participants gave their preferences a second time for the other five scenarios, just as in Studies 2a and 2b, and then answered a demographic survey. We predicted that there would be an effect of the explicit/implicit condition for the preference changes associated with the SQO and SO hypotheticals, such that participants would shift significantly toward preferring computers in the implicit condition but not in the explicit condition.

Participants were again recruited from MTurk. Data was collected from 154 participants (76 female).

Results

Collapsed across implicit/explicit condition, the changes in preference associated with each intervention were significant (for SQO, p=.017; for SO, p<.001; for B, p<.001).

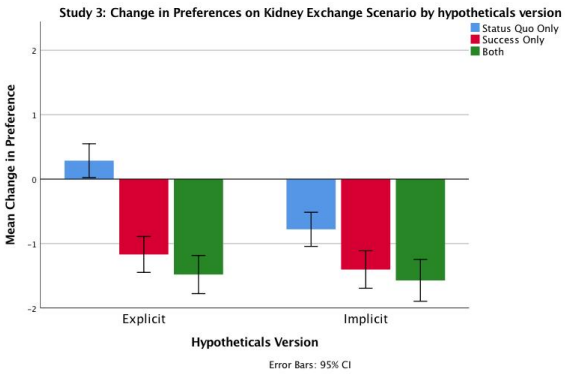

Figure 4: Changes in preference separated by hypothetical condition in Study 3.

Furthermore, a series of paired t-tests shows that the effect of each intervention was distinct from the others, providing us a ranking of the interventions by strength: B is the most powerful intervention, followed by SO, followed by SQO well behind. An ANOVA of dependent variables by implicit/explicit condition shows that the distinction in types of hypotheticals only made a difference for the SQO intervention (p<.001; other two interventions p>.25). In fact, the implicit/explicit distinction created a bidirectional effect for the SQO intervention: subjects who saw the explicit version actually shifted 0.29 points toward humans, rather than computers (p=.033), while those who saw the implicit version shifted their preference 0.78 points toward computers (p<.001). This difference, and the lack of difference for other hypothetical preferences, is shown in Figure 4. In sum, what we see is that for the SQO intervention, making the lack of evaluative information explicit recasts the non-evaluative information as evidence against computers' abilities.

Conclusion and Discussion

Some of the most interesting results of our investigation are null results. From Study 1, it is surprising that no psychological measures consistently tracked trends in preference. One would think that at least some of these measures—perhaps risk sensitivity or risk seeking (RPS), owing to the weighty nature of these decision-making roles, or empathy (AEQ), since it seems an apt candidate for predicting preferences—but nothing was significant. Trivially, one might expect older participants to be warier of such a new and potentially intrusive technology, but even age effects were nowhere to be found. The only aspects that consistently made a difference were how good participants judged the two types of agents to be at solving complicated problems and a self-report of prior experience with these agents making these kinds of decisions. The exposure result, one might think, connects to a well-studied effect in social psychology, the mere-exposure effect. As relayed by Robert Zajonc (2001), its modern progenitor, the effect describes a phenomenon by which the exposure to some item—a word, a foodstuff, a song—generates the gradual emergence of a preference

for the item. If this is what is happening in the case of our subjects' preferences for AI systems making decisions, then this would fit nicely in a well-defined story. However, we do and we do not see this. In fact, the pattern of effects we see in our interventions in Study 2 is quite strange. The N intervention, which includes no substantive information about computers making decisions in kidney exchange scenarios, only mentioning that they could make such a decision, significantly shifts preferences in the direction of computers. This seems like a mere-exposure effect; the N intervention *merely exposes* subjects to the notion that computers could do this task, and suddenly subjects express (more of) a preference for computers to do it. However, the B intervention consistently produced a larger effect than the N intervention, so mere exposure cannot be doing all of the work. Moreover, the SQO and SO interventions, which provided more information that might encourage a subject to develop a preference for computers (prevalence of implementation and positive evaluations, respectively) in fact generated less of a change in preference than the N intervention. We also saw some generalization from the kidney exchange intervention into another medical domain, but the effect was greatly reduced and which intervention generated it was inconsistent.

The results of Studies 2a and 2b, in which providing no evidence convinced people to change their mind but providing some evidence did not, demands consideration and explanation. Our hypothesis, tested in Study 3, was that at least some participants who were given just some evidence were noticing the absence of other relevant evidence and inferring, on these grounds, the lack of other supporting evidence, and therefore not shifting their preferences. In other words, participants in the SO and SQO conditions noticed that something was missing, and this caused them to be skeptical of the scant evidence they were given. To some degree, this was borne out by the data from Study 3: when we made that absence obvious, even to those who would not have inferred the absence, the skepticism it generated actually caused a reactive shift toward preference for humans. Meanwhile, when that absence was left unspoken, there was a significant preference shift toward computers, though of a smaller magnitude than for the other hypothetical versions of the kidney exchange scenario. This was only the case for the SQO hypothetical, not the SO hypothetical. We have two potential explanations. One is that if participants only receive a mention of current deployment, they may take this as evidence of success as well, thinking that these machines would not be deployed if they did not work well. However, when presented with the possibility of direct evidence of success (as in the within-subjects design of Study 3) rather than inferential evidence, status quo evidence ceases to stand in as a measure of success. On this interpretation, the exposure effect is still mostly due to an increased perception of efficacy and success. Along similar lines, mentioning another possible evaluative criterion may cause subjects to think about it when they never would have otherwise. After all, if someone tells you, "There is no poison in this wine," that may decrease your preference for the wine, because you are now thinking about poison. A similar sort of phenomenon may explain the reactive move toward preferring humans on the explicit SQO hypothetical.

In any case, the matter of how participants are interpreting this evidence and the manner in which they use it to update their preferences is still a puzzle. One (perhaps trivial) hypothesis is that different people may have different cognitive profiles that cause them to interpret and apply the information they learn in different ways. Our initial reading of the results of Study 3 points in this direction. We set on a level field those who could infer the absence of complementary evidence (and thereby reframe what information the evidence conveys) and those who could not by making the absence explicit, but we did not do anything to figure out what the difference is between those two types of subjects. To flesh out this hypothesis and pursue it further would require an idea of what psychological construct would fill this interpretative role, which was beyond the immediate scope of this project.

All in all, though, we were able to extract relatively stable preferences in dilemmas that are rapidly becoming more prevalent in our everyday lives, and we were able to successfully shift these preferences. Importantly, we showed that these preferences are contingent not on values that a particular person holds but rather previous experience with computer agents acting in these ways. This suggests that, as computers continue to be implemented in roles that carry more and more consequential weight, and as their implementation becomes more visible, this might in itself generate acceptance of the phenomenon.

As AI becomes further integrated into decision-making roles in our various social institutions, understanding whether (and why) the public or consumers may be made uneasy or reassured by knowing an AI system is at the reins is important, as is understanding how these reactions and preferences may be changed through public outreach and informational campaigns. We have provided some preliminary, exploratory results for the "whether" and "how," at least in one domain. It is our hope that in the future, this research may be extended to other domains, and perhaps by comparing between social domains, we can further understand not only the preferences that the public holds in this important matter, but the deeper psychological processes that explain how we develop and update these preferences in response to rapid social and technological change.

Acknowledgments

The authors thank the Future of Life Institute for funding this research. The authors are, of course, solely responsible for the content.

References

Aquino, K., and Reed II, A. 2002. The self-importance of moral identity. *Journal of Personality and Social Psychology* 83(6):1423–1440.

Arnett, J. 1994. Sensation seeking: A new conceptualization and a new scale. *Personality and Individual Differences* 16(2):289 – 296.

Beadle, J. N.; Sheehan, A. H.; Dahlben, B.; and Gutchess, A. H. 2015. Aging, empathy, and prosociality. *The Journals of Gerontology: Series B* 70(2):213–222.

Bonnefon, J.-F.; Shariff, A.; and Rahwan, I. 2016. The social dilemma of autonomous vehicles. *Science* 352(6293):1573–1576.

Coutlee, C. G.; Politzer, C. S.; Hoyle, R. H.; and Huettel, S. A. 2014. An abbreviated impulsiveness scale constructed through confirmatory factor analysis of the barratt impulsiveness scale version 11. *Archives of Scientific Psychology* 2(1):1–12.

Davis, R.; Libon, D. J.; Au, R.; Pitman, D.; and Penney, D. L. 2015. Think: inferring cognitive status from subtle behaviors. *AI Magazine* 36(3):49–60.

Dickerson, J., and Sandholm, T. 2015. Futurematch: Combining human value judgments and machine learning to match in dynamic environments. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 622–628.

Everett, J. A. C. 2013. The 12 item social and economic conservatism scale (secs). *PLOS ONE* 8(12):1–11.

Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review* 5(5-15).

Freedman, R.; Schaich Borg, J.; Sinnott-Armstrong, W.; Dickerson, J.; and Conitzer, V. 2018. Adapting a kidney exchange algorithm to align with human values. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Graham, J.; Nosek, B. A.; Haidt, J.; Iyer, R.; Koleva, S.; and Ditto, P. H. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology* 101(2):366–385.

Haidt, J.; McCauley, C.; and Rozin, P. 1994. Individual differences in sensitivity to disgust: A scale sampling seven domains of disgust elicitors. *Personality and Individual Differences* 16(5):701 – 713.

Hare, R. D.; Harpur, T. J.; and Hemphill, J. F. 1989. Scoring pamphlet for the self-report psychopathy scale: Srp-ii.

Meertens, R. M., and Lion, R. 2008. Measuring an individual's tendency to take risks: The risk propensity scale1. *Journal of Applied Social Psychology* 38(6):1506–1520.

Noothigattu, R.; Gaikwad, S. N. S.; Awad, E.; D'Souza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. D. 2018. A voting-based system for ethical decision making. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Wallach, W., and Allen, C. 2009. *Moral machines: teaching robots right from wrong*. Oxford University Press.

Zajonc, R. B. 2001. Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science* 10(6):224–228.