When Do Humans Heed AI Agents' Advice? When Should They?

Richard E. Dunning, Baruch Fischhoff and **Alex L. Davis**, Carnegie Mellon University, Pittsburgh, PA, USA

Objective: We manipulate the presence, skill, and display of artificial intelligence (AI) recommendations in a strategy game to measure their effect on users' performance.

Background: Many applications of AI require humans and AI agents to make decisions collaboratively. Success depends on how appropriately humans rely on the AI agent. We demonstrate an evaluation method for a platform that uses neural network agents of varying skill levels for the simple strategic game of Connect Four.

Methods: We report results from a 2 × 3 between-subjects factorial experiment that varies the format of AI recommendations (categorical or probabilistic) and the AI agent's amount of training (low, medium, or high). On each round of 10 games, participants proposed a move, saw the AI agent's recommendations, and then moved.

Results: Participants' performance improved with a highly skilled agent, but quickly plateaued, as they relied uncritically on the agent. Participants relied too little on lower skilled agents. The display format had no effect on users' skill or choices.

Conclusions: The value of these AI agents depended on their skill level and users' ability to extract lessons from their advice.

Application: Organizations employing AI decision support systems must consider behavioral aspects of the human-agent team. We demonstrate an approach to evaluating competing designs and assessing their performance.

Keywords: artificial intelligence, advisory systems, decision making under uncertainty, trust in AI, team performance

INTRODUCTION

Artificial intelligence (AI) technologies are being used in an increasingly wide variety of tasks, performing as well as humans on tasks as diverse as navigating the London underground (Gibney, 2016), engaging in conversational speech (Xiong et al., 2017), and extracting information from natural language (Narasimhan, Yala, & Barzilay, 2016). Many of these advances come from deep learning, which has been used in scientific discovery (Gilmer et al., 2017; Green et al., 2022), medical science (Rajkomar, Dean, & Kohane, 2019), and strategic decision making (Brown & Sandholm, 2017; Vinyals et al., 2019). These breakthroughs have raised hopes that AI technologies can support human decision making in high-stakes environments, such as criminal justice sentencing, employment and hiring, and healthcare (Albert, 2019; Gifford, 2018; Rajkomar et al., 2019). They have also raised concerns that these technologies may do more harm than good without appropriate human collaboration and supervision (e.g., Buolamwini & Gebru, 2018; Mitchell et al., 2019; Future of Life Institute, 2023). Humans have essential roles in supervising programs that work correctly most of the time, but need human intervention when they fail (Endsley, 2017). That role can be particularly difficult with programs whose high reliability induces an inappropriate sense of security (Green 2021). Thus, the success of AI depends on keeping humans appropriately in the loop (Chiou & Lee, 2023; Endsley, 2017, DSB, 2016). Here, we demonstrate a general method for evaluating such efforts.

How human operators interact with and trust automation is a long-standing topic in human factors research (Lee & See, 2004; Meyer & Lee, 2013; Parasuraman, 2000; Parasuraman & Riley, 1997). Recent applications with AI suggest that operators often have difficulty evaluating and implementing system advice, leading to suboptimal performance (McNeese, Demir, Cooke, & Myers, 2018; Bartlett & McCarley, 2017; Dzindolet et al., 2000). In some cases, performance may be worse with an aid than without (Alberdi, Povyakolo, Strigini, & Ayton,

Address correspondence to Richard E. Dunning, Department of Engineering and Public Policy, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA; e-mail: rdunning@andrew.cmu.edu

HUMAN FACTORS
Vol. 0, No. 0, ■■ ■, pp. 1-14
DOI:10.1177/00187208231190459
Article reuse guidelines: sagepub.com/journals-permissions

2004). Unless users know when and how much to trust an aid, they may use it too much, too little, or inappropriately (Aoki, 2020; Gao and Waechter, 2017, Lee and See, 2004). In their seminal work, Lee and See (2004) defined trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Lee and See (2004) identified purpose, process, and performance as critical to that trust. Many other factors such as workload and training have been found to affect human use of automation and AI, but system reliability and performance have the greatest overall impact (Hancock et al., 2011, Kaplan et al., 2021).

One potential obstacle to achieving appropriate trust with AI algorithms built on deep learning, like AlphaZero, is that they are a black box (Hassoun, 2003) with an opacity that makes it difficult for users to create mental models of system processes. Unable to explain their rationale, these algorithms depend on user trust built through interaction. To improve team performance, developers' attempt to increase the transparency of algorithms by providing additional information about their internal workings, so that user can develop more accurate mental models (Bansal et al., 2019).

One potential resource for increasing transparency is the probability of success for possible actions that some algorithms calculate as part of their computational framework (Silver et al., 2018). Compared to a categorical best choice, those probability distributions might provide useful information for assessing trust. Studies have found that humans value receiving numerical probabilities and can often use them effectively, when they are associated with well-defined events (Erev & Cohen, 1990; Gaube et al., 2021; Lipkus, 2007; Zhang et al., 2020). Probabilities may have little value, though, if a system has such great predictive value that humans just defer to it or such little predictive value that humans stop using it (Meyer, 2004; Parasuraman & Riley, 1997; Wickens & Dixon, 2007). Probabilities may also have little value when systems impose such great cognitive load that users cannot attend to the probabilities (Peters et al., 2006; Wickens, 2008) or where they recommend infeasible actions (Bertuccelli & Cummings, 2011).

We offer a general method for assessing how successful people are at deciding when to rely on AI-based advisors, illustrated with a realistic, engaging task. The task provides human users with repeated trials that allow learning about their own abilities, the abilities of the AI aid, and the opportunities for human-aid team collaboration. Our task is patterned after the two-alternative forced-choice (2AFC) tasks often used with 'yes/no' independent stimuli (e.g., alarms, color discrimination, target detection) (e.g., Wiczorek & Meyer, 2019; Bartlett & McCarley, 2017), applying Signal Detection Theory to those responses (Green & Swets, 1966). We extend that paradigm to an n-alternative forced-choice (nAFC) task in a strategy game with advice from AI agents whose abilities players must learn from observed performance. The task is the game Connect Four, chosen because it is complex enough that an AI aid could be useful, with an algorithm that can provide success probabilities for future moves, but is simple enough not to impose cognitive load that will keep users from attending to the probabilities.

We examine two psychological processes key to the success of AI systems: how much human operators trust them and how much they learn from them. We manipulated two potential determinants of trust: (a) how skilled the system is, as revealed in the course of play and (b) how it expresses its recommendations, as a categorical best move or a probability distribution over the set of possible moves.

HUMAN SUBJECTS STUDY

The present research created and demonstrated a platform for studying human-AI collaboration. It used the simple strategic game Connect Four and AI agents with varying skill levels. Connect Four players alternately place yellow or red discs in one of seven columns of a 6 × 7 grid, alternating with another player. Players win by getting four straight discs in a single row, column, or diagonal (Hasbro, 2009). The game is an adversarial, zero-sum, sequential, perfect information. It was first solved by computer in 1988 by James Allen (Allen, 2010). It was chosen because it is easy enough for someone with no prior experience to learn to play during the experiment, but hard enough that humans cannot easily compute its solution.

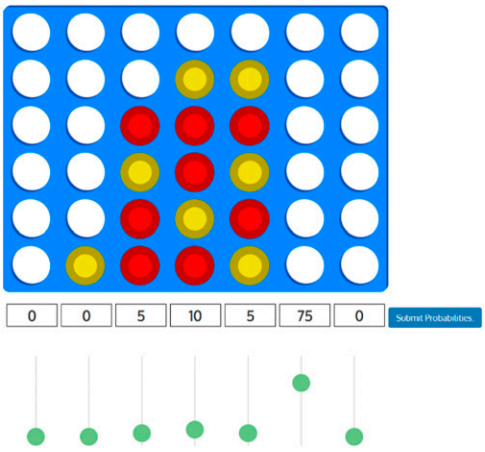

Figure 1. Example Connect Four board, with probability sliders below each column.

Because Connect Four belongs to the class of deterministic games that reinforcement learning methods can solve to any desired degree of mastery (Schrittwieser et al., 2020), we were able to create agents with different ability levels, by varying the amount of model training. Our implementation of AlphaZero was trained by Anthony Young (2018) and adapted here for stimuli modification, skill assessment, and data collection (see supplemental material for more agent details). As seen in Figure 1, these agents have 1 (low), 5 (mid), or 20 (high) cycles of neural network training. In the same way, we created an opponent, with 3 cycles of training, roughly equivalent to human performance in pretest trials, and an oracle, with 50 cycles, used to define the best possible solution. Table 1 shows each agent, its cycles of training, and its error compared to an optimal Connect Four solver during development (Young, 2018). The skill level is the percentage of rounds that each Agent agreed with the oracle during our platform development (see Table 1). Our method uses these measures to evaluate the agents, the human, and the human-agent team (explained below).

Participants played 10 games, with random assignment to a control group or one of the six conditions in a 2 × 3 between-subjects design, with the factors of *display* (categorical, probabilistic) and AI *agent skill level* (high, medium, low). The control group received no AI recommendations and simply played Connect Four 10 times. In each round, players made a *provisional* move, saw the agent's recommendation, and then made a decision to use their provisional move or switch to the agent's recommendation, producing the human-agent *team* move. The design was extensively pretested to reduce the chance that the interface affected results (e.g., players ignoring the probability distribution because the display was hard to understand) (Elliott et al., 2012).

TABLE 1: Artificial Intelligence (AI) Agent Characteristics

Agent	Training Cycles	Error Rate	Skill
Low	1	16%	39%
Opponent	3	9%	59%
Mid	5	6%	69%
High	20	2%	81%
Oracle	50	<1%	100%

Hypotheses

For ease of exposition, we formulated our hypotheses in directional form, adopting the perspective of a proponent of AI agents. We hypothesized that:

H1: Human skill will be greater when players receive a probability distribution rather than a categorical recommendation.

H2: Human skill will improve over trials, improving more with more skilled AI agents.

H3: Human skill will be greater with any AI agent than with none (control group).

H4: Team skill will be greater than human skill.

H5: When players and AI agents disagree, players will more often defer to agents with higher skill levels.

We were, a priori, agnostic about these hypotheses, as there were plausible reasons favoring and opposing each. Regarding H1, humans may or may not be able to use the additional probabilistic information, depending on how well they can interpret its content and manage the additional

cognitive load. Regarding H2, players may benefit from any AI agent or only from agents with demonstrably superior ability. Regarding H3, players may or may not be able to learn from observing the agent and trying to extract usable lessons from the algorithm, whose calculations reflect patterns that humans cannot see. Regarding H4, humans may or may not mistakenly defer to an underperforming agent. Regarding H5, humans may or may not be able to evaluate the quality of agent advice well enough to rely more on better agents, in these complex tasks. This study was preregistered with Open Science Foundation (https://osf.io/wzecx/?view_only=ae2be15e074040c98b0ea998b0058b5e).

METHODS

Participants

We recruited 156 participants through Amazon's Mechanical Turk, using a link to a website that hosted the experiment. Participants were paid a base rate of Pennsylvania's minimum wage ($7.50/hr) for completing the experiment, along with a bonus that depended on their proportion of games won (=$4.50 × proportion of wins). The bonus was meant to provide an incentive to optimize the decision, rather than just complete it quickly (Young, 1967). Participants were randomly assigned to one of the seven groups (six experimental and one control), with approximately 20 participants in each treatment group (see Figure 3).

Stimuli

For each move, participants were shown the current board state, as in Figure 2. Participants had 1 to 7 columns into which chips could be placed depending on whether columns were already filled. Participants played the yellow chips, while the AI opponent, which always moved first, played red. The board states were dynamic, based on game play.

Figure 2 shows the displays for participants who received probabilistic (top) or categorical (bottom) recommendations. Both displays used the same calculation, with the categorical display indicating the column with the highest probability.

Task

Before making each move, humans were asked to evaluate the board and then make a provisional move, which was compared to an oracle to calculate *human skill*. They then indicated the probability that each possible move was the best, using vertical sliders or typing a percentage for each column. Although pretest players were able to use the probability response interface, few players in the experiment gave probabilities other than 100% for more than a few plays. As a result, we did not analyze the probability responses, as originally planned. Participants with an AI agent then received its recommendation (as a probability distribution or categorical best move), which was used to calculate *agent skill*. Participants then made their final choice, which we used to calculate *team skill*. In cases where the provisional move and agent recommendation disagreed, we recorded whether that player switched to the agent's recommendation and whether that change was to a better or worse move, as defined by the oracle. After the player moved, the website updated the game board and the opponent moved again. Game play proceeded until one side won or the game ended in a draw.

Measures

The following measures were computed for each player, for each move in each game.

- Human Skill: The proportion of moves where the human's initial choice, before seeing the AI recommendation, matched the oracle's (near-optimal) choice.
- Agent Skill: The proportion of moves where the AI agent's recommendation matched the oracle's choice.
- Team Skill: The proportion of moves where the team choice, made by the human after seeing the AI agent's recommendation, matched the oracle's choice.

The following measures were computed for players with agents, for moves where the agent's

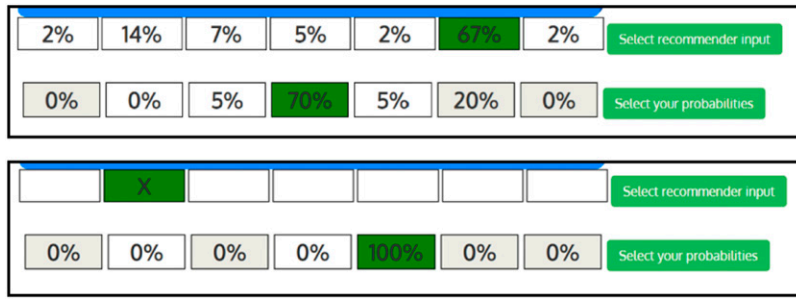

Figure 2. AI agent recommendation for the probability distribution display (top) and the categorical display (bottom). In each display, the top row is the AI agent's recommendation. The bottom row is the human player's probability distribution for each possible move winning, made before receiving the AI agent's recommendation.

recommended move disagreed with the player's provisional move.

- Appropriate Acceptance: The proportion of instances where the player switched to an agent's recommendation that agreed with the oracle.
- Appropriate Rejection: The proportion of instances where the player did not switch to an agent's recommendation that disagreed with the oracle.

Design

Figure 3 depicts the 2 (recommendation display) × 3 (agent skill) between-subjects design, along with the control group. The agent gave either a probability distribution, for each possible move being best, or a categorical recommendation, of the move with the highest probability. Participants in the AI groups were randomly assigned to agents with low, medium, or high skill. Before playing, participants were introduced to the Connect Four game and the interface. Participants played 10 games with a brief survey between games, followed by a longer survey at the end of the study.

Procedure

The experiment lasted approximately 45 minutes, including consent, tutorial and instructions, game play, and final surveys. All participants completed their tasks remotely over the Internet, using their preferred computer web browser. After each game, participants were shown their number of wins, losses, and draws. After the 10 games, they were asked for demographic information, thanked, and given a code to secure compensation from Amazon Mechanical Turk.

This research was approved by the Institutional Review Board at Carnegie Mellon University as "Human-AI Interaction Experiment." All procedures were performed in accordance with NIH Office for Human Research Protections regulations and in compliance with relevant laws and institutional guidelines. Informed consent was obtained from each participant.

Beta Regression

While general linear modals can be powerful tools for data interpretation and analysis, three main assumptions must hold the following: (1) normally distributed residuals, (2) homogeneous residual variance, and (3) residuals independent of each other (Graybill, 1961; Seber, 1966; Sokal & Rohlf, 1969). Our skill measures are double-bounded [0%, 100%], potentially violating the normality assumption (Ferrari & Cribari-Neto, 2004; Verkuilen & Smithson, 2012). Beta regression, developed by Kieschnick and McCullough (2003) and Ferrari and Cribari-Neto (2004), assumes that the dependent variable has a beta distribution with respect to linear predictors, making it appropriate for double-bounded dependent measures, such as rates, proportions, and percentages (Cribari-Neto and Zeileis, 2010). Cribari-Neto and Zeileis (2010) developed the most popular R

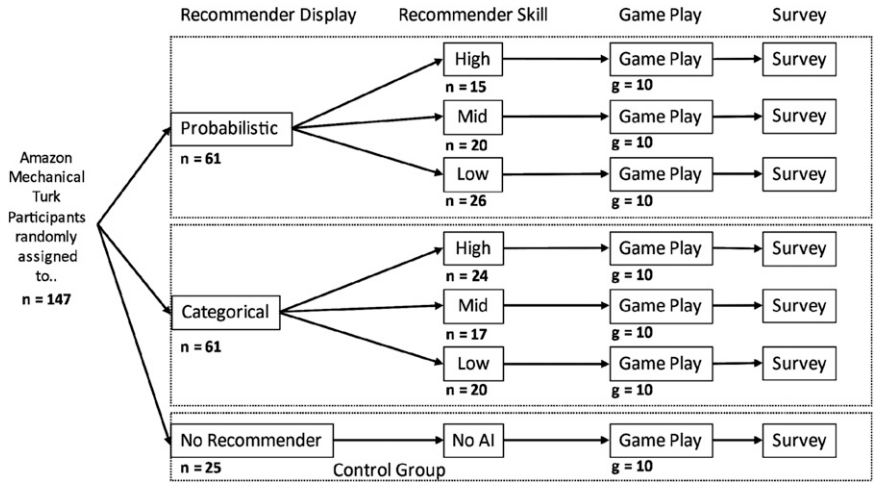

Figure 3. Experimental design. Subjects (n) were randomly assigned to experimental groups to play 10 games (g = 10) with agents in their treatment conditions.

package betareq. More recently, Brooks et al. (2017) developed the R package glmmTMB, which allows for mixed-effects models, hence is better suited to our experiment which has multiple measures from each subject (Brooks et al., 2017). To accommodate values at 0 and 1, we adopted a transformation proposed by Smithson and Verkuilen (2006); $n^{-1}(y(n-1)+0.5)$, where n is the sample size. We chose the logit link to connect the modeled statistic with the regressors so that both are unbounded (McCullah & Nelder, 1989).

RESULTS

On each move, in each game, human skill is scored as 1, if the human's highest probability, before seeing the AI agent's recommendation, matches that of the oracle; or as 0, if the two do not match. The AI agent's skill and subsequent team skill were scored similarly. Agent skill varied by game, depending on the human's play (e.g., it was higher with weaker human players). As a result, we treated it as a continuous variable, whose mean was expected to approximate that observed in the training (Table 1). During preanalysis, 10 participants were removed from the data for apparently noncompliant response patterns. Some repetitively lost games with minimal moves and minimal time; indicating that they were attempting to complete the experiment quickly; 1 apparently used an online solver, with 9 games of identical extremely high scores. Figure 1 shows the remaining 147 participants by their assigned group. Because our control group had no agent, we used a mixed effects 2 × 3 factorial design for the treatment conditions, which we compared separately to the control (Marini, 2003). Our mixed effects analyses accounted for varying subject intercepts and skill change over game play. To compare the treatment and control groups, we used a single factor with 7 levels (experimental groups). To determine statistical significance, we applied Chi-squared Wald Type III tests and contrasted results using Tukey–Kramer pairwise comparisons.

Human Skill Changes Through Game Play

Figure 4 shows human skill performance across game play for the control group, with each point representing the proportion of matches (to the oracle's play) for each player for that game. The skill level of players in the control group, with no AI recommender, did not increase or decrease significantly, over the games ($p = .324$) (see Table A1). The group's mean skill across all games was 0.397 (SD = 0.067), which resulted in winning 12.8% of the games.

Figure 5 shows human skill across game play for the six treatment groups, with each point representing the proportion of matches (to the oracle) for each human for that game. There was

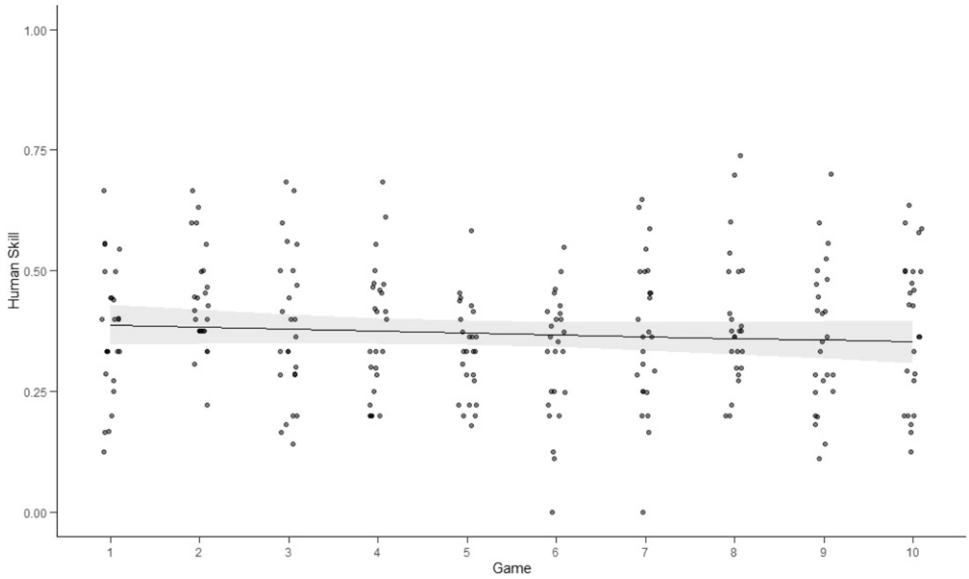

Figure 4. Human skill scores for each player, as a function of game play, for the control group (with no AI aid). Points are proportions correct for individual players of each game. Curves reflect Beta regressions with 95% CI.

a significant interaction between game play and AI agent skill ($X^2(1, N = 122) = 12.004$, $p < .001$; details in Tables A2 and A2-1). However, display type (categorical vs. probabilistic) had no significant impact on human skill ($p = .946$) nor any significant interaction with AI skill or game play ($p = .901$, $p = .807$, respectively; Table A2). Tukey Pairwise comparisons (Table A2-5) found greater improvement in human skill across games when players were paired with an AI agent of middle or high skill level, compared to pairing with a low skill agent ($p < .001$, $p = .005$, Cohen's $d < .05$, respectively). Players paired with middle and high skill agents improved at a similar rate ($p = .840$).

A mixed effects beta regression, treating the 7 experimental groups as levels in one experimental group factor, found no overall difference in mean human skill (see Table A3-2); however, there was a significant interaction effect of experimental group and game ($X^2(6, N = 147) = 26.921$, $p < .001$; see Table A3-1). The only statistically significant differences were that humans viewing the categorical display with a high skill agent improved more than those in the control group, with no display ($t = -3.444$, $p = .011$, Cohen's $d = .05$) and those with a probabilistic display and a low skill agent improved less than those with a categorical display with a high skill agent ($t = -3.602$, $p = .006$, Cohen's $d = .05$) or a middle skill agent ($t = -3.074$, $p = .035$). See Table A3-3 and Supplementary Materials for additional detail.

Resolution of Human-Agent Disagreement

Overall, the agent agreed with the oracle, while the human did not, on 23.9% of plays (SD = 0.15). In those cases, humans appropriately accepted the agent's recommendation 46.1% of the time (SD = 0.36). The human matched the oracle, but the agent did not, on 11.1% of plays (SD = 0.11). In those cases, humans appropriately rejected the agent's recommendation 59.8% of the time (SD = 0.38).

There was no difference between the two displays in the appropriate acceptance rate ($p = .948$) nor any significant interaction between display and agent skill level or game (Table A4). Subjects paired with more reliable agents appropriately accepted recommendations at a higher rate ($t = -4.283$, $p <$

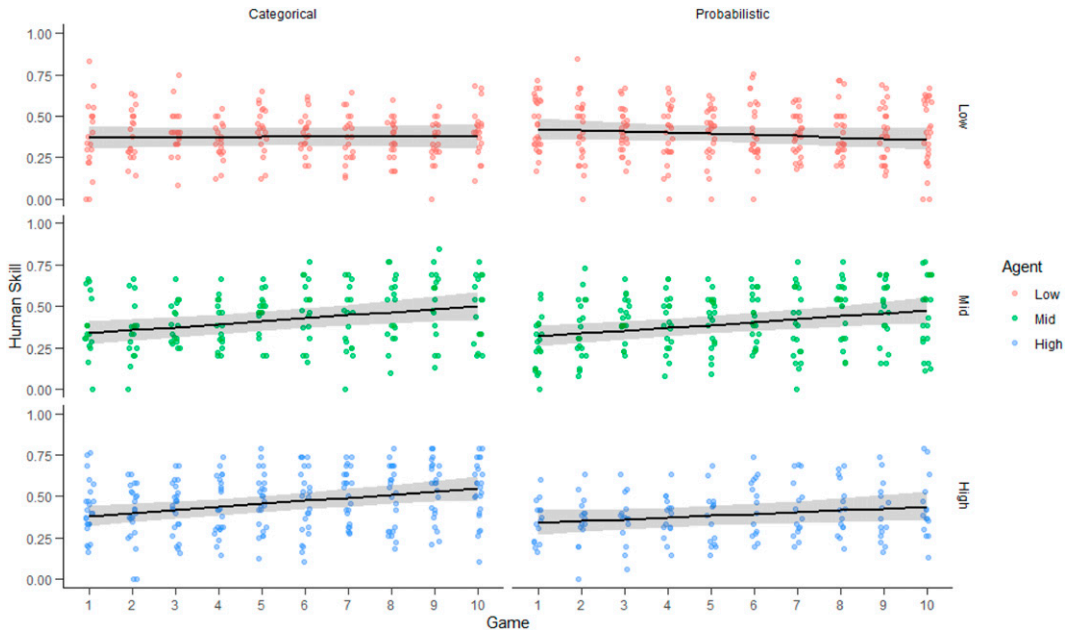

Figure 5. Human skill as a function of game play, for low (top), mid (middle), and high (bottom) agent skill levels and for categorical (left) and continuous (right) displays. Points are proportions correct for individual players of each game. Curves reflect Beta regressions with 95% CI.

.001, Cohen's d = .52, Table A4-3) and acceptance increased across game play with more skilled agents (Table A4-4). With the probabilistic display, teams with high skill agents increased their rate of acceptance compared to low skill teams (Table A4-5). The supplemental material and Figure (S)M5 have details. There was no difference between the two displays in the appropriate rejection rate, nor any interaction between display and agent skill level or game (Table A5). Appropriate rejection was more likely with less skilled agents ($X^2(1, N = 122) = 6.478$, $p = .011$, Table A5-1); however, the rate did not change with game play. The Supplemental Material and Figure SM3 have details.

Team Skill

Figure 6 shows the performance of the human-agent team, as reflected in team skill scores for players' final moves, made after seeing the agent's recommendations. Here, too, the display made no difference; nor was there an interaction between display and agent skill level or game (Table A6). Pooling the display groups, AI skill largely determined team skill ($X^2(1, N = 122) = 18.871$, $p < .001$, Tables A6 and A6-1). Teams with middle and high skill agents had higher scores than teams with low skill agents across all games ($t = -6.406$, $p < .001$, Cohen's d = .44, Table A6-3), and increased over game play (Table A6-5). The team outperformed the low skill agent on 60.4% of plays (SD = 0.49), the middle skill agent on 3.2% (SD = 0.17), and the high skill agent on only 0.24% (SD = 0.049). Team skill was higher than human skill for 71.4% of plays with the categorical display (SD = 0.46) and for 67.2% of plays (SD = 0.47) with the probabilistic display. Figure (S)M7 compares the treatment groups by agent and display.

Figure 7 compares human, agent, and team skill across game play. Overall team skill was significantly determined by experimental group and the interaction of group and game play ($X^2=(6, N = 147) = 20.555$, $p = .002$; see Table A7-1). Teams with high and middle skill agents showed greater skill and skill improvement over game play, compared to the control group, while those with low skill agents did not (see Tables A7-2 and A7-3).

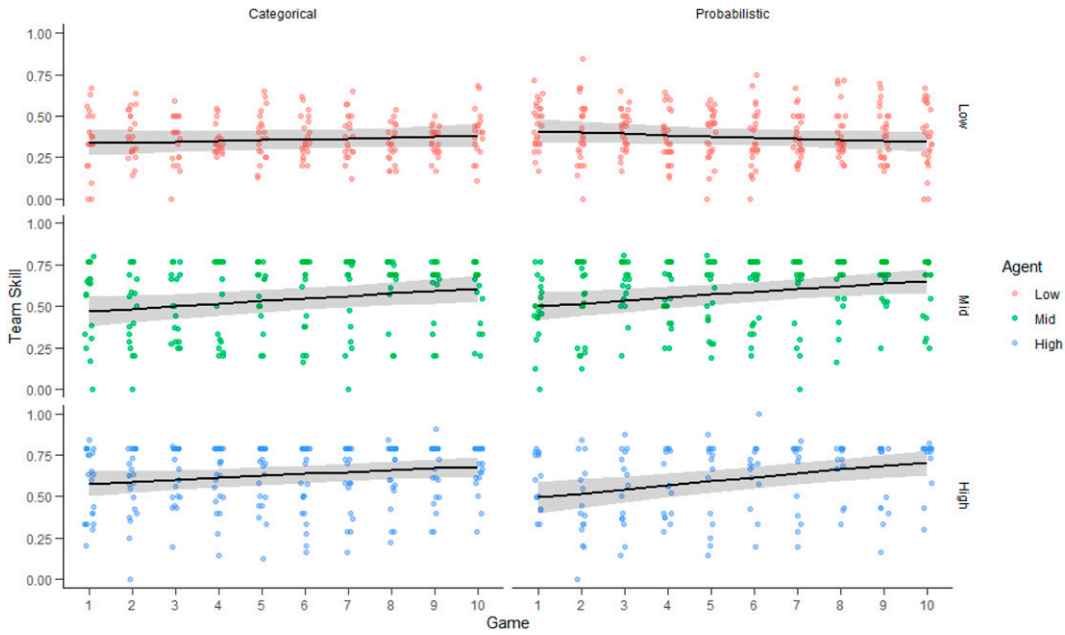

Figure 6. Mean team skill, as a function of game play, for agent skill levels low (top), middle (middle), and high (bottom) and for categorical (left) and continuous (right) displays. Points are proportions correct for teams of each game. Curves reflect Beta regressions with 95% CI.

Win Rates

The scores reported above describe the details of human, agent, and team play. The ultimate, aggregate performance measure is how many games are won and lost. Control group humans won 1.28 (SD = 1.54) games, on average. Teams with high and middle skill agents won more games that teams with low skill agents or control humans (two sample *t* test, t = 3.124, $p < .001$, Cohen's D >1.0), which won games at the same rate (t = 1.337, $p = .091$, Cohen's D <0.3). Display made no difference except that with low skill agents, teams won more games with the probabilistic display than with the categorical display (t = 2.298, $p = .011$, Cohen's D = 0.513).

DISCUSSION

The risks and benefits of adopting AI technology in decision making will depend on how well humans understand the strengths and weaknesses of AI aids, so that they afford them appropriate trust. The present study demonstrated a platform for studying human-AI team performance, using the simple strategic game of Connect Four. It illustrated the research platform with one possible advising system, where the human proposed a move, received an agent recommendation, and then made the move. We compared two possible displays of agent recommendations: probabilistic (distribution over possible moves) and categorical (the best move, as implied by that distribution). We also varied agent skill, as a function of the number of training cycles for the underlying deep reinforcement learning algorithm. The opponent was always an algorithm whose performance matched that of the average human in pretests.

In terms of our hypotheses, we found that:

H1: Overall, participants performed similarly with the two displays, seemingly unable to use the additional information provided by the probabilistic display. The few significant display differences had no obvious pattern and seem best attributed to chance.

H2: Over the course of play, human skill improved slightly with the middle and high skill AI agents. Human skill did not

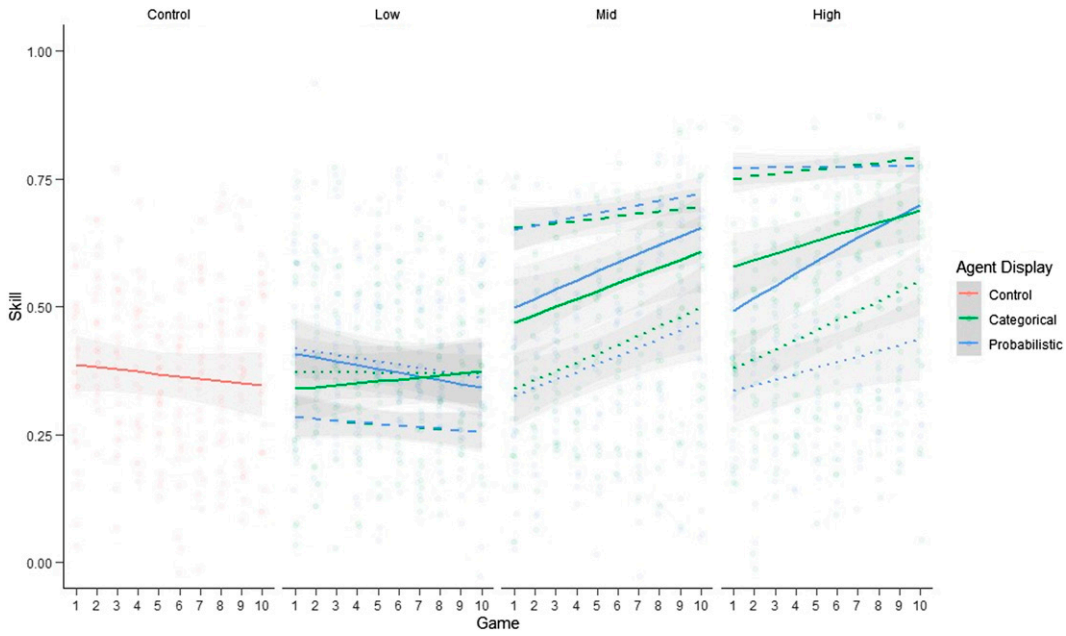

Figure 7. Human, agent, and team skill as a function of game play, for control (left), low (left-center), middle (center-right), and high (right) skill agents and for categorical (left) and probabilistic (right) displays. Points are proportions correct for subjects of each game. Curves reflect Beta regressions with 95% CI for human subjects (dotted lines), agents (dashed lines), and teams (solid lines).

improve for either the control group or humans paired with low skill agents.

H3: Human skill was not better for players when paired agents compared with the control for most treatment groups. Players who were paired with high skill agents and received the categorical display had higher human skill than control condition players.

H4: Team skill was greater than agent skill for 60% of the games with the low skill agent, but rarely exceeded that of the middle or high skill agents. These results might be promising for using high skill AI systems, problematic for less skilled ones.

H5: Over the course of play, human players were increasingly able to reject inappropriate recommendations from low skill agents and accept appropriate recommendations from middle and high skill agents. However, they still resolved many disagreements inappropriately. They tended to defer uncritically to more skilled agents, while ignoring good advice from low skilled agents, a pattern seen elsewhere (Bartlett & McCarley, 2017; Lee & See, 2004; Wiczorek & Meyer, 2019). As a result, agent skill was a powerful predictor of team performance.

As mentioned, human skill was similar with both displays in almost all analyses. The one significant interaction found that, for the probabilistic display, humans appropriately accepted high skilled agent recommendations over game play at a higher rate than with low agents. A speculative account is that the probabilistic display helped players to see the limits to low skill agents' advice, while strengthening faith in the high skill agents. Overall, though, the cognitive load of processing the probabilistic display may have counterbalanced the value of the additional information that it provided (Endsley, 2017).

The cognitive load of producing probabilities presumably accounts for why participants rarely provided responses other than 100%, for the probability of their move being the best one (Erev & Cohen, 1990). As a result, we did not conduct planned analyses of the appropriateness of participants' confidence in their moves (calibration) and their ability to use the additional information in the display to protect against overconfidence (Dzindolet et al., 2002; Peters et al., 2006).

Whether changes in performance like those observed here would warrant investment in an AI system would depend on the application. In some settings, marginal improvements are highly valuable; in others, less so. The answer would also depend on the AI system's capital costs (acquisition, installation, upgrades), operating costs (training time, fatigue), opportunity costs (competing investments of money and personnel), and system costs (deskilling, habituation). Finally, it would depend on how well human-agent conflicts are identified and resolved. Our scoring treated all games and moves as equally important. That will not always be the case.

Although our system underwent extensive user testing, there was little evidence of human learning over the 10 rounds of play in the control group, without a skilled AI aid. Players' deferral to the high skill agent suggests that they learned little about its strategies, other than to trust them. Such learning is essential when agents are imperfect, hence may need to be overruled, or when agents are not available (e.g., a computer malfunction takes an aid offline or an attack disables AI functions in a field of combat). Such learning is, of course, essential when practice with an AI agent is intended as training.

Limitations

The present results reflect the performance of participants recruited through Amazon Mechanical Turk. They were paid relatively well for the MTurk world, with an average hourly rate of $12.10, structured to include a performance incentive. They also worked on a relatively engaging task. Nonetheless, their performance and sensitivity to task features (e.g., the display) may have been less than that in real-world settings. In future research, time-stamp data might provide a useful predictor of cognitive effort, performance, and sensitivity.

Although the task used here was a simple game, it has some key properties of many AI decision support systems. The human user faces an unfamiliar advisor, with unknown skill, and unexplained recommendations. The human must decide whether to trust the AI advisor's recommendations, both when they support and when they contradict the human's intuitive ones. In life, users might have to wait for feedback, if they receive it at all. In the present task environment, where they received immediate feedback, participants were able to learn something about how much to trust their AI agent, but little about how to proceed on their own. The opacity of the AI agent, even with the probability display, did not allow users to improve their play. Although our application used a deep reinforcement learning aid, the behavioral issues should be similar with other AI decision aids. Those responsible for deploying such systems must evaluate the short- and long-term effects of introducing them. The present study offers a testing protocol and metrics that could guide those evaluations.

CONCLUSIONS

As organizations evaluate AI technologies intended to aid decision making, they must consider behavioral aspects of the human-agent team. Policies on the procurement, maintenance, and operational deployment of AI decision systems should specify their requirements for how well humans can tell when to trust the AI agent, resolve disagreements, and learn over time. We found that humans had some imperfect ability to tell how much to trust these AI agents and learn their skill level, in the complex, deterministic environment of the task used here. That ability improved with the middle and high skill agents, but decreased with the low skill agent—which itself decreased over the course of play. Whether that pattern recurs with other tasks and environments is an empirical question, as is what can be done to improve human learning about and from AI agents.

ACKNOWLEDGMENTS

Partial funding for this research was provided by The Swedish Foundation for the Humanities and Social Sciences under its research program for Science and Proven Experience. We thank Mengda (Martin) Liu, Jia Shen, and Jinghua Huang, for designing the experiment website, and Anthony Young, for sharing his original website code and training our agents.

KEY POINTS

- For a complex strategy task, AI recommendations improved task performance with skilled AI agents, but not with unskilled ones.
- Human players appeared to follow the recommendations of highly skilled AI agents uncritically.
- Human players did not benefit from the additional information in AI agent recommendations expressed as a probability distribution over possible options, rather than as a best-guess categorical recommendation.

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

Alberdi, E., Povyakalo, A., Strigini, L., & Ayton, P. (2004). Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology*, *11*, 909–918. https://doi.org/10.1016/j.acra.2004.05.012

Albert, E. T. (2019). AI in talent acquisition: A review of AI-applications used in recruitment and selection. *Strategic HR Review*, *18*(5), 215–221. https://doi.org/10.1108/shr-04-2019-0024

Allen, J. D. (2010). *The Complete Book of Connect 4: History, Strategy, Puzzles*. Puzzle Wright Press.

Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. *Govern. Inf. Q*, *37*(4), 101490. https://doi.org/10.1016/j.giq.2020.101490

Bansal, G., Nushi, B., Kamar, E., Lasecki, W., Weld, D., & Horvitz, E. (2019) (In press). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *The Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP-19)*.

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking Aided Decision Making in a Signal Detection Task. *Human Factors*, *59*(6), 881–900. https://doi.org/10.1177/0018720817700258

Bertuccelli, L. F., & Cummings, M. L. (2011). Scenario-based robust scheduling for collaborative human-UAV visual search tasks. In 50th IEEE Conference on Decision and Control and European Control Conference, pp. 5702–5707. https://doi.org/10.1109/CDC.2011.6160994

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielson, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400.

Brown, N., & Sandholm, T. (2017). Libratus: The superhuman AI for no-limit poker. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), pp. 5226–5228.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency, pp. 77–91.http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Chiou, E. K., & Lee, J. D. (2023). Trusting automation: Designing for responsivity and resilience. *Human Factors*, *65*(1), 137–165. https://doi.org/10.1177/00187208211009995

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, *34*, 1–24.

Defense Science Board. (2016). Report of the defense science board summer study on autonomy. *Autonomous Weapon Systems: An Exploration of Issues and Recommendations* (pp. 41–178). Nova Science Publishers, Inc.

Dzindolet, M. T., Beck, H. P., & Pierce, L. G. (2000). Encouraging human operators to appropriately rely on automated decision aids. In Proceedings of the 2000 Command and Control Research and Technology Symposium. International Command and Control Institute.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Endsley, M. (2017). From Here to Autonomy: Lessons Learned from Human-Automation Research. *Human Factors*, *59*(1), 5–27. https://doi.org/10.1177/0018720816681350

Elliott, L. R., Jansen, C., Redden, E. S., & Pettitt, R. A. (2012). *Robotic telepresence: perception, performance, and user experience*. Army Research Laboratory (US). Aberdeen Proving Ground (MD) (Report No.: Arl-Tr-5928.

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes*, *46*, 1–18. https://doi.org/10.1016/0749-5978%2890%2990002-Q

Ferrari, S. L. P., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*, 799–815. https://doi.org/10.1080/0266476042000214501

Future of Life Institute (2023). Pause open AI experiments: An open letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Gao, L., & Waechter, K. A. (2017). Examining the role of initial trust in user adoption of mobile payment services: an empirical investigation. *Inf. Syst. Front*, *19*(3), 525–548. https://doi.org/10.1007/s10796-015-9611-0

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S., Lermer, E., Coughlin, J., Guttag, J., Colak, E., & Ghassemi, M. (2021) (In press). Do as AI Say: Susceptibility in Deployment of Clinical Decision-Aids. *npj Digital Medicine*, *4*(31). https://doi.org/10.1038/s41746-021-00385-9

Gibney, E. (2016). Google's AI reasons its way around the London Underground. *Nature*. https://doi.org/10.1038/nature.2016.20784

Gifford, R. (2018). Legal technology: Criminal justice algorithms: AI in the courtroom. *The Proctor*, *38*(1), 32–33.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. In 34th International Conference on Machine Learning, ICML 2017, pp. 2053–2070.

Graybill, F. A. (1961). An Introduction to Linear Statistical Models (Vol 1). McGraw-Hill.

Green, A. G., Yoon, C. H., Chen, M. L., Ektefaie, Y., Fina, M., Freschi, L., Groschel, M. I., Kohane, I., Beam, A., & Farhat, M. (2022). A convolutional neural network highlights mutations relevant to antimicrobial resistance in mycobacterium tuberculosis. *Nature Communications*, *13*. Article 3817.https://doi.org/10.1038/s41467-022-31236-0

Green, B. (2021). The flaws of policies requiring human oversight of government algorithms. *Social Science Research Network*. (SSRN Scholarly Paper ID 3921216).https://doi.org/10.2139/ssrn.3921216

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Krieger.

Hasbro (2009). Connect four instructions. https://www.hasbro.com/common/documents/dad2614d1c4311ddbd0b0800200c9a66/1EF6874419B9F36910222EB9858E8CB8.pdf

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*, 517–527. https://doi.org/10.1177/0018720811417254

Hassoun, M. (2003). *Fundamentals of Artificial Neural Networks*. MIT Press.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2021). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, *00*(0), 1–25. https://doi.org/10.1177/00187208211013988

Kieschnick, R., & McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. *Statistical Modeling*, *3*, 193–213. https://doi.org/10.1191/1471082X03st053oa

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors. *The Journal of the Human Factors and Ergonomics Society*, *46*, 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making*, *27*(5), 696–713. https://doi.org/10.1177/0272989x07307271

Marini, R. (2003). Approaches to analyzing experiments with factorial arrangements of treatments plus other treatments. *HortScience*, *38*(1), 117–120. https://doi.org/10.21273/HORTSCI.38.1.117

McCullah, P., & Nelder, J. A. (1989) (In press). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming with a synthetic teammate: insights into human-autonomy teaming. *Human Factors*, *60*(2), 262–273. https://doi.org/10.1177/0018720817743223

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors Summer*, *46*(2), 196–204. https://doi.org/10.1518/hfes.46.2.196.37335

Meyer, J., & Lee, D. (2013). *Trust, Reliance, and Compliance, The Oxford Handbook of Cognitive Engineering* (pp. 109–124). Oxford Library of Psychology, Oxford University Press. Chapter 6.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., & Gebru, T. (2019). Model cards for model reporting. In Conference on Fairness, Accountability, and Transparency (pp. 220–229). Association for Computing machinery. https://doi.org/10.1145/3287560.3287596

Narasimhan, K., Yala, A., & Barzilay, R. (2016). Improving information extraction by acquiring external evidence with reinforcement learning. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 2355–2365). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D16-1261

Parasuraman, R. (2000). Designing automation for human use: Empirical studies and quantitative models. *Ergonomics*, *43*, 931–951. https://doi.org/10.1080/001401300409125

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*(5), 407–413. https://doi.org/10.1111/j.1467-9280.2006.01720.x

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2020) (In press). Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *arXiv preprint arXiv:1911.08265v2*.

Seber, G. A. F. (1966). *The linear model and hypotheses*. Springer. Springer Series in Statistics.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(No. 1), 54–71. https://doi.org/10.1037/1082-989X.11.1.54

Sokal, R. R., & Rohlf, F. J. (1969). *Biometry: The principles and practices of statistics in biological research* (1st ed.). W. H. Freeman and Company.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., & Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*. https://doi.org/10.1038/s41586-019-1724-z

Verkuilen, J., & Smithson, M. (2012). Mixed and Mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics*, *37*(1), 82–113. https://doi.org/10.3102/1076998610396895

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*, 449–455. https://doi.org/10.1518/001872008X288394

Wickens, C. D., & Dixon, S. R. (2007) (In press). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212. https://doi.org/10.1080/14639220500370105

Wiczorek, R., & Meyer, J. (2019). Effects of trust, self-confidence, and feedback on the use of decision automation. *Frontiers in Psychology*, *10*(MAR), 1–12. https://doi.org/10.3389/fpsyg.2019.00519

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., & Zweig, G. (2017). Achieving human parity in conversational speech recognition. Microsoft. Report: MSR-TR-2016-71.https://doi.org/10.48550/arXiv.1610.05256

Young, A. (2018). *AZFour: Connect four powered by the alphazero algorithm*.

Young, F. W. (1967). Twelve-choice probability learning with payoffs. *Psychonomic Science*, *7*(10), 353–354. https://doi.org/10.3758/BF03331120

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 295–305. https://doi.org/10.1145/3351095.3372852

Richard E. Dunning received a BS in mechanical engineering from the UW-Madison, an MS in engineering management from UW-Platteville, and an MS in systems engineering from Southern Methodist University. He is a PhD student in the Department of Engineering and Public Policy at Carnegie Mellon University.

Baruch Fischhoff is a professor in the Department of Engineering and Public Policy and Institute for Politics and Strategy, Carnegie Mellon University. He studies decision making, with a focus on empowering people to participate actively in public and private decisions. He went to the Detroit Public Schools, Wayne State University (mathematics, psychology), and the Hebrew University of Jerusalem (psychology). He is an elected member of the National Academy of Sciences and of the National Academy of Medicine. His books include *Acceptable Risk*, *Risk: A Very Short Introduction*, *Risk Communications: The Mental Models Approach*, and *Counting Civilian Casualties*.

Alex Davis is an associate professor in the Department of Engineering and Public Policy, Carnegie Mellon University. He studies decision making with a focus on statistical modeling. He is a graduate of Northern Arizona University (BS in psychology), and Carnegie Mellon University (PhD in behavioral decision making).

Date received: November 18, 2022
Date accepted: June 20, 2023