frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 12 March 2025
DOI 10.3389/fcomp.2025.1405436

OPEN ACCESS

EDITED BY
Ilaria Tiddi,
VU Amsterdam, Netherlands

REVIEWED BY
Tayana Soukup,
Imperial College London, United Kingdom
Kendall Carmody,
Florida Institute of Technology, United States

*CORRESPONDENCE
Nikolos Gurney
gurney@ict.usc.edu

RECEIVED 22 March 2024
ACCEPTED 26 February 2025
PUBLISHED 12 March 2025

CITATION
Gurney N, Pynadath DV and Miller JH (2025) Willingness to work as a predictor of human-agent team success.
Front. Comput. Sci. 7:1405436.
doi: 10.3389/fcomp.2025.1405436

Willingness to work as a predictor of human-agent team success

Nikolos Gurney[1]*, David V. Pynadath[2] and John H. Miller[3]

[1]Institute for Creative Technologies, Viterbi School of Engineering, University of Southern California, Playa Vista, CA, United States, [2]The Ken Kennedy Institute, Rice University, Houston, TX, United States, [3]Department of Social and Decision Sciences, Dietrich College of Humanities and Social Sciences, Carnegie Mellon University, Pittsburgh, PA, United States

Research shows that the effectiveness of human-agent teams depends heavily on human team members' prior experiences, whether from direct teaming activities or relevant domain knowledge. While researchers have proposed various mechanisms to explain this relationship, we present a simpler alternative explanation: experience serves primarily as an indicator of a person's fundamental willingness to engage in teaming tasks. We introduce a measure called "willingness to work" that quantifies this underlying disposition. Our empirical analysis demonstrates that this straightforward metric robustly predicts human-agent team performance. Beyond its practical value as a predictive tool, this reconceptualization of the experience-performance relationship necessitates a fresh examination of existing findings in the field. The results suggest that a team member's basic willingness to invest effort may be more fundamental to success than previously recognized mechanisms.

KEYWORDS

human-agent teaming, human-AI interaction, human-computer interaction, hybrid intelligence, user modeling, human-machine integration

1 Introduction

Extensive research tests the hypothesis that a person's experience in a task domain influences the success of a new human-agent team (HAT) (Huang and Bashir, 2017; Demir et al., 2018; McNeese et al., 2018; O'Neill et al., 2022). Empirical evidence suggests that this effect holds for experience that comes from simple domain-relevant settings (Chen et al., 2011) to past teaming with AI-agents (Hafizoğlu and Sen, 2018; Gurney et al., 2023c). Researchers have devoted considerable efforts to explaining the mechanisms through which experience improves these teaming outcomes. Prominent examples of this are the widely studied (and debated) tandem hypotheses that experience can impact trust in automation and trust in automation predicts teaming outcomes (Lee and See, 2004; Hancock et al., 2011; Hoff and Bashir, 2015; Huang and Bashir, 2017; Lewis et al., 2018). Although these relationships undoubtedly explain some variance in teaming outcomes (see Huang and Bashir, 2017 for a study of how interaction dynamics impact trust), they likely do not explain all of the variance, because like all measures, experience alone has its shortcomings. For example, what constitutes relevant experience might differ significantly across domains—experience as a pilot in a human-autonomy flying team is markedly different than experience in a gaming environment (Demir et al., 2018; Pynadath et al., 2023). Moreover, experience alone does not equate to willingness, or motivation, to do a task. Work to untangle how experience and motivation are related abound in the management literature, with one prominent model arguing that considering the valence of an experience (whether it is positive or negative) is crucial to predicting its impact

on motivation (Seo et al., 2004). In applied settings, however, it is not always feasible to dissect and measure a worker's motivations nor is it necessary. All that is needed is a simple measure of how willing the worker is to *work*. Although experience is important, and undeniably linked to this willingness, it does not entirely capture it.

We propose a simple alternative explanation of why experience is frequently predictive of teaming outcomes: experience serves as a proxy for an individual's willingness to do the task, i.e., their willingness to work (WTW). If a person has experience in a domain and is returning to the same for more work, then there is some validation of their willingness to do that specific task (work). People unwilling to do a given task will not return for more (outside of coercion). Of the people that do return, however, one can expect natural variation in their WTW. Based on this idea, we hypothesized that for a given population of experienced workers, the variation in their WTW, as measured by previous effort, will predict the amount of future effort they are willing to invest.

Extensive research explores an individual's willingness to work (or interact) *with* AI [e.g., robots (You and Robert, 2018; Paluch et al., 2021; Verma and Singh, 2022), intelligent virtual agents (Sycara and Lewis, 2004; Cafaro et al., 2016; Boukaram et al., 2021), etc.] and the impact that such willingness has on teaming success. Here we consider the simpler hypothesis that an individual's basic WTW drives many such results. We demonstrate the viability of this hypothesis using data from a study of human heuristics and biases during complex choices when teaming with an AI. Critically, the experiment includes a solo-effort baseline in which participants completed two versions of the task before teaming with the AI helper. We find that including a person's effort from their solo work as an independent variable in teaming models significantly improves their accuracy and predictive abilities. This simple and often costless metric was positively correlated with human-AI team performance. Moreover, when such data are available, an agent can readily model this relationship and use the model to improve teaming outcomes.

Definition: *Willingness to Work* (WTW) is a person's fundamental quality or state of being disposed to engage in an effortful task. This fundamental state is primordial to higher-order concerns such as morals, individualism, gratification, compensation etc. Although a person who is industrious or has good work ethic will likely have a high WTW, the opposite is not necessarily true. That is, a person high in WTW will not necessarily be recognized as industrious or of good work ethic. All else being equal, a person higher in WTW will be more likely to invest effort in a given task than a person who is lower in WTW.

Our approach is fundamentally different from prior approaches in that we view WTW as a fundamental state of an individual that predicts other outcomes. Prior work has focused on the accompaniment aspect of HATs—that is, how working in a team with an AI agent impacts a person's effort or other measurable outcome. Our measure reverses the hypothesis and posits that for a given task, each individual has some basic WTW that impacts the outcome, regardless of the HAT setting.

2 Materials and methods

We analyze data from an experimental study of HATs completing a dial-tuning task that manipulated the presence of overt anchors (informing the human team members of the best possible outcome in a given task) and whether the HAT was in a loss or gain frame (attempting to not lose points vs. attempting to gain points). Both anchoring (Tversky and Kahneman, 1974; Chapman and Johnson, 1999; Epley and Gilovich, 2006) and framing (Tversky and Kahneman, 1974, 1985; Kahneman, 2011) are correlated with dramatic impacts on human judgment and decision making. Anchoring is simply the tendency to use salient information as a reference in decision making. Framing describes how casting a decision in a positive vs. negative way can alter decision outcomes. Complete descriptions of the task, experiment, and results are reported in the papers from which we drew the data (Gurney et al., 2022a, 2023a,b). Each participant completed the task four times: twice on their own and then twice with the help of an AI agent, in which case they ceded control of one of the dials to the agent. Performance was incentivized with bonus payments for uncovering better dial settings in each task. Participants were free to break off the search whenever they chose.

2.1 Task summary

Participants used two on-screen dials to search an unseen landscape for its highest location (see Figure 1). Conceptually, these dials moved participants in perpendicular directions in the landscape (one left-right and the other up-down). Each landscape was a unique, algorithmically drawn, constrained environment containing 576 locations that ranged in elevation on the interval $[0, 32]$. The locations were arranged on a 24×24 grid that was continuous over the edges, meaning that moving off of the left (top) edge of the map led to the right (bottom) edge of the map and vice versa. Participants were unaware of the landscape-search aspect of the task. Instead, they were only asked to optimize the dial settings and find what they believed to be the highest value. For both the solo and HAT landscapes, participants completed, in a random order, a simple (single-peaked) and a complex (four-peaked) landscape. The peaks in the landscapes always took a value of 32 units; the other three peaks in the complex landscape were randomly selected without replacement from the interval $[26, 31]$. Importantly, the dials were interdependent and the landscapes were nonlinear.

Team search was a turn-taking endeavor: participants first adjusted the dial that they controlled and submitted it for evaluation, then the AI agent took that information and decided whether or not to adjust its dial. A stochastic model based on a simulated annealing algorithm guided the AI helper's decision making. Although checking every possible setting would take a participant more than 20 min, the AI could do it in fractions of a second. To facilitate richer interactions, the AI was limited to deciding whether to look near or far from its current location, after which a setting was randomly selected and either accepted or rejected given the output from the simulated annealing algorithm. Participants then received feedback on what the AI decided and the value of the new setting.

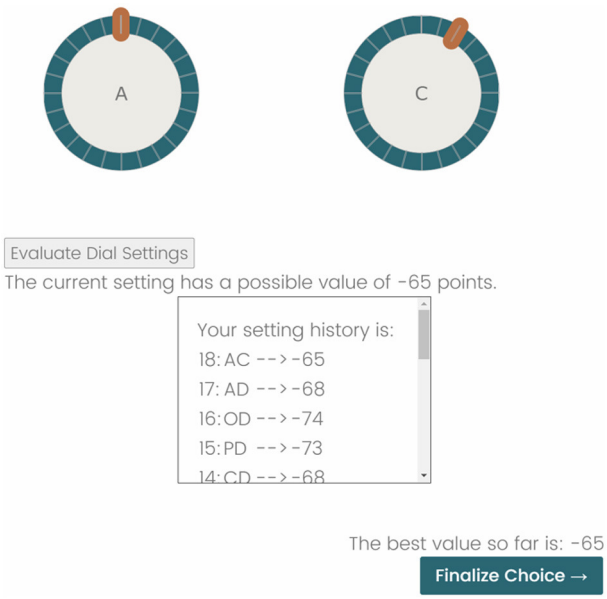

FIGURE 1
A screenshot of the experimental interface. In this example, the participant is in a loss frame (note the negative points; −65 is better than −68) and there is not an anchor value (if it were available, it would be indicated above the dials). Adjusting the dials entails clicking and dragging the orange handle around the turquoise ring. Clicking the *Evaluate Dial Settings* button retrieves the associated landscape elevation. These are tracked in the setting history box. When the participant believes that they have found the best (or a good enough) dial setting, clicking the *Finalize Choice* button advances them to the next task.

2.2 Experiment summary

The experiment crossed an explicit anchoring treatment (participants were informed of the best they could do on a given landscape vs. not) with a framing treatment (the incentivized goal was to avoid losing as few points vs. gain as many points as possible) resulting in a 2×2 design. To control for scaling effects, the landscape values were randomly perturbed such that the loss frame values were in the $[-100, 0]$ interval and the gain frame values were in the $[0, 100]$ interval. This meant that participants in the no-anchor conditions had to discover what was a "good" landscape value for each task, i.e. a prior effort was not informative of a current effort. The solo efforts were always completed before the team effort and the order of the simple and complex landscapes was randomized in both instances. After completing the tasks, participants answered a set of questions about their own performance, the performance of the AI, and their related opinions. We briefly summarize the experimental results below; full results are available in Gurney et al. (2023a,b).

2.3 Dial tuning experimental results summary

Participants' solo effort outcomes are available in Gurney et al. (2023a). The experiments were designed such that simple linear approaches (regression, ANOVA, etc.) could readily model the outcome measures of interest (e.g., highest discovered location). 398 participants from Prolific Academic completed the experiment, 172 participants identified as male, 218 as female, and eight as other. The average participant age was 32 years. Two hundred three participants indicated that they were college graduates with a four-year degree or higher. We considered one participant as an outlier: they evaluated 607 settings for one of their solo tasks and 581 for one of their HAT tasks, more than the complete set of combinations (576). This participant achieved a perfect solo score and a near perfect HAT score. They did not find the global maximum for the one peaked landscape in the HAT, although they submitted 96 fewer dial settings in the HAT than when they worked alone on the one peaked landscape (18 vs. 114). They also submitted This was also three times the effort of the next most ambitious participant. Given that this behavior would have an outsized impact on any analysis, we decided to remove this participant from the data set, leaving 397 observations.

Participants did better, as expected, on the single-peaked than the four-peak landscapes during the solo effort. Moreover, doing the single-peak landscape first was correlated with better outcomes on the subsequent four-peak landscape, but not vice-versa. Anchoring and framing yielded main effects, but not interaction effects: being in a loss frame was correlated with a longer average search duration (measured as the number of dial submissions), and having the explicit anchor was correlated with a shorter average search duration, all else being equal. Also, participants in the loss frame committed more effort to fine-tuning the dial settings than participants in the gain frame. Lastly, the anchoring treatment did not meaningfully alter the participants' search strategies.

The HATs' outcomes are available in Gurney et al. (2023b). On average, HAT scores were worse than solo effort scores, a result driven, primarily, by significantly shorter search duration from the loss framing participants. HATs also spent more time exploring distal locations in the landscape than fine-tuning the dials to uncover local topographies, i.e. exploiting local knowledge. Nevertheless, HATs in the loss frame conditions fared significantly better than those in the gain frame. No matter the treatment condition, the HATs did worse on the four-peak than on single-peak landscapes. Interestingly, participants exerted less effort in each successive task, however, the only significant decrease in effort between any two tasks happened when they joined a HAT. Again, participants in the loss frame made the largest adjustments and suffered the biggest setbacks for it.

Critically, participants' effort, as measured by the number of submissions made, in the solo tasks was correlated with their effort during the team tasks. This insight sparked our hypothesis and serves as the foundation for our WTW models.

3 Willingness to work models

We define three WTW metrics: willingness to work *solo* (WTWs), willingness to work in a human-agent *team* (WTWt), and a measure of a person's willingness to work in a HAT *relative* to alone (WTWr). WTWs is the total number of dial settings submitted by a participant across their solo tasks. WTWt is the

number of dial settings a participant submitted during their first HAT task. WTWr is WTWs minus the number of dial settings submitted by the HAT divided by the total of submissions. In other words, WTWr is a person's willingness to work in the team in relation or proportion to their general willingness to work. We use WTWs and WTWr to predict the total team score of a HAT and WTWt to predict the second team score. The score on a given task is simply the highest elevation uncovered. The maximum available points in a given task was always 32 (the highest elevation), and the scores are simply the fraction of points earned. The total team score and the second team score are percentage values. For example, if a hypothetical team earned 20 points on the first HAT task and 28 on the second, then their total team score would be $\frac{20+28}{32\times2} = 0.75$, or 75%, and their second team score $\frac{28}{32} = 0.875$, or 87.5%. Our models also include control variables for the treatment conditions. The findings reported in Gurney et al. (2023a,b) support main effects for the treatment conditions, but not an interaction, thus we do not include an interaction term for the treatment conditions in our models.

We rely on two classes of models, which we chose for their ability to capture the theorized effects without introducing unnecessary complexity. The first models are ordinary least squares (OLS or multiple linear) regression models in which the score of interest (total team score or second team score) is predicted by the appropriate WTW measure and the treatment condition variables. The basic model is:

$$S_i = \beta_0 + \beta_1 W_i + \beta_2 A_i + \beta_3 F_i + \epsilon_i \quad (1)$$

Where S represents the relevant score measure, W the relevant WTW measure, A the anchoring condition of a participant, and F the framing condition of a participant. β_0 is the constant (intercept) value, which does not have much meaning for this study since continuous independent variables are anchored at zero for ordinary least squares regression models and because participants had to submit at least one dial setting per task. The β_1 values in the WTWs and WTWt models indicate the expected change in a score for each additional dial setting submitted. Since there are no interactions in these models, the effect of a given WTWs or WTWt value is computed by simply multiplying it by the appropriate β_1.

The observed range of the dependent variables is censored for the linear models (scores could only range from 0 to 100%). This feature creates the possibility of over or underfitting and violating normality assumptions. One solution for this is to fit a more complicated censored regression model, commonly known as a tobit, which we did and include in the Supplementary material. These models suggest the same interpretation of the data, although result in larger coefficient values. Since we are not interested in exact point estimates, we decided to rely on the more parsimonious models.

The relative willingness to work measure allows us to understand whether a participant's general WTW interacts with their WTW in the HAT. The model we report for WTWr differs from the basic paradigm of Equation 1 in that it interacts WTWs and WTWr. Here, we anticipate WTWs to have an effect when WTWr is zero, which occurred when a participant exerted the same effort on their own as in the HAT. WTWr, however, should not be significant, as it would represent a case when WTWs was zero, which was not possible. Significant results for WTWs and the interaction term WTWs × WTWr will suggest that (1) a participant's raw willingness to work and willingness to work in the HAT are predictive of HAT success and (2) they share some correlation.

The second class of models includes ordinal logistic regressions, for which we sort the total and second scores into three classes by splitting the distribution into terciles for low, medium, and high achievement. Terciles were selected based on the amount of data available: more cuts would undermine the statistical power of the models, which bifurcating the distribution provided less insight into the phenomenon. These models serve an illustrative purpose, as it is often more convenient to categorize outcomes than make exact point estimates. Recall that for such models, Y is an ordinal outcome with J categories. $P(Y \leq j)$ is the cumulative probability of Y less than or equal to a given category $j = 1, ..., J-1$. The log odds, or logit, is simply the log of the cumulative probability divided by the $P(Y > j)$. Since $P(Y > j)$ is 0, the log odds reduce to $logit(P(Y \leq j))$ and our basic model of the terciles is:

$$logit(P(S_i \leq j)) = \beta_0 - \eta_1 W_i - \eta_2 A_i - \eta_3 F_i - \epsilon_i \quad (2)$$

$S, W, A, and F$ serve the same abbreviation function as in Equation 1. β in these models indexes the intercept values while η indexes the coefficient values. The former coefficients simply indicate where the team score variables were cut to make the terciles and are generally not used in the interpretation of ordinal model results, thus we omit them from the results. The range of the submission counts for the terciles for the total team score of Equation 2 are [4, 143], [144, 198], and [199, 337], and for the second team score of Equation 2 they are [2, 53], [54, 88], and [89, 124]. The cuts for scores were 53 and 77% for the WTWs model and 78 and 90% for the WTWt model. These cuts fit with the analyses in Gurney et al. (2023b) which suggest that the HATs did better on their second than first tasks. Since the ordinal models serve only an illustrative purpose and are not part of our hypothesis testing, we forgo reporting the WTWr model for brevity's sake.

4 Empirical strategy

We first fit the above models using the complete data set. We compare the OLS models to a control model using F-tests and the ordinal logistic models to a control model using χ^2-tests. In these tests, the control model is the null hypothesis against which the alternative is tested. A significant result for either the F or χ^2-test means that adding the WTW measures is justified based on the amount of variance in the data that the richer model explains relative to the control model.

To gain further insight into the predictive value of WTW, i.e., when it might perform poorly, we conducted five-fold cross-validation for the WTWs model depicted in Equation 1 using the *caret* package in R (Kuhn, 2008). We bootstrapped this process ($n = 1,000$) to get an expected prediction value for every observation in the data set and stored the predicted values for each test case. We then computed the difference between the true values and the bootstrapped predictions by simply subtracting the

latter from the former. The true values are left-skewed (participants tended to find locations better than the average task value), a feature that the new statistic will reflect. Because of this, z-scores are not meaningful. As an alternative, we simply analyze the 5% of observations on either side of the distribution, i.e., outliers with extreme predictions. The same in-depth look at the predictive abilities of WTWt is possible, but largely redundant, as is a similar effort of cross-validation for the categorical models. We forgo these for brevity.

5 Results

5.1 WTWs and WTWt multiple linear regression models

The outcomes measures, the total score as a percentage achieved by the HAT across both tasks [HAT Task 1 + 2, column (1) of Table 1] and for the second task [HAT Task 2, column (2) of Table 1], lend themselves to linear modeling. Our models predict these values using either a participant's total submissions during their solo effort (willingness to work solo, WTWs) or from the first HAT task plus controls for the treatment conditions (willingness to work team, WTWt), respectively.

The overall WTWs linear regression model showed statistical significance [$R^2 = 0.256$, $F_{(3,393)} = 45.043$, $p < 0.001$] and WTWs significantly predicted the total HAT score ($\beta = 0.002$, $p < 0.001$). In other words, submitting one additional dial setting during the solo effort predicted a 0.18% higher total HAT score according to the WTWs model. Moreover, we reject the null hypothesis that including WTWs does not contribute to the model with just the treatment controls ($F = 124.759$, $p < 0.001$).

The overall WTWt linear regression model was also statistically significant [$R^2 = 0.160$, $F_{(3,393)} = 24.991$, $p < 0.001$], although it did explain less variance. WTWt significantly predicted HAT Task 2 ($\beta = 0.004$, $p < 0.001$). Submitting one additional dial setting during the first HAT effort predicted a 0.45% higher HAT score during the second task according to the WTWt model. Moreover, we can again reject the null hypothesis that including WTWt does not contribute to the model with just the treatment controls ($F = 69.429$, $p < 0.001$).

In both models, the amount of a HAT score predicted by WTW is relatively modest. The median number of submitted dial settings during the solo effort was 33. The distribution was heavily right-skewed; the maximum number of submissions was 337, and 282 participants submitted 50 or fewer dial settings. In exploratory analyses, we trimmed the data to only include efforts in which a participant submitted from 10 to 50 dial settings, which left 208 observations for the WTWs model and 192 for the WTWt model. The models were still statistically significant [$R^2 = 0.243$, $F_{(3,204)} = 21.771$, $p < 0.001$; $R^2 = 0.098$, $F_{(3,188)} = 6.810$, $p < 0.001$] and, interestingly, although the predicted effect of both WTWs and WTWt did change, the changes were modest (WTWs increased to 0.45% while WTWt dropped to 0.37%).

TABLE 1 Multiple linear regression models.

Variable	Dependent variable (score)	
	HAT Task 1 and 2	HAT Task 2
	(1)	(2)
WTWs |WTWt	0.002***	0.004***
	(0.0002)	(0.001)
Constant	0.720***	0.730***
	(0.013)	(0.016)
Treatment controls	Yes	Yes
Observations	397	397
R^2 | Adjusted R^2	0.256 | 0.250	0.160 | 0.154
RSE (df = 393)	0.128	0.164
F Stat. (df = 3; 393)	45.043***	24.991***

*p < 0.05; **p < 0.01; ***p < 0.001.

TABLE 2 Five-fold cross validation resampling results.

Statistic	N	Mean	SD	Min	Max
RMSE	5	0.128	0.016	0.112	0.154
R-squared	5	0.271	0.050	0.221	0.334
MAE	5	0.098	0.013	0.086	0.116

5.2 Five-fold cross-validation of WTWs multiple linear regression model

We used a five-fold cross-validation method to evaluate the WTWs model presented in Equation 1. The sample sizes of the training sets were 317 or 318 observations. The absolute difference between the predictions of the model and the observations was low, suggesting good performance ($MAE = 0.098$), and the correlation between the predictions made by the model and the actual observations reflected that of the model reported in column (1) of Table 2 ($R^2 = 0.271$). These are considerable improvements over a similar cross-validation of the null model that does not have the WTWs measure ($MAE = 0.114$, $R^2 = 0.029$).

Figure 2 plots the results of the bootstrapping effort: the difference in the predicted mean scores and the observed scores of the HATs using five-fold cross-validation. As anticipated, the skew in the distribution reflects a similar skew in the HAT scores. The first insight from analyzing the outlier data is that the no anchor, gain-framing treatment condition ($n = 15$) was over-represented in the outlier data. Meanwhile, the anchor, loss-framing treatment condition was under-represented ($n = 5$). Given the relatively small size of the outliers sample ($n = 40$), we are hesitant to make definitive claims about the data. However, a logistic regression model with the no anchor, gain-framing treatment condition as the reference condition predicting whether a given observation is in the outlier group did suggest that a HAT with a participant in the anchor, loss-framing treatment condition was about 33% less likely to be in the outlier group.

A two-sample t-test comparing the HAT score of the outliers from the left and right side of the distribution of differences (the

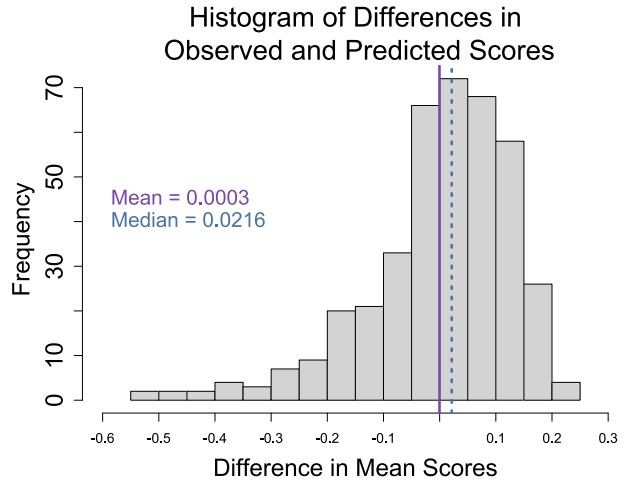

FIGURE 2
Histogram of the differences in predicted and observed scores. Scores for each HAT were computed but adding the highest achieved values from both efforts and dividing by the maximum score possible. The difference was calculated by subtracting the predicted mean scores from the observed scores. A negative (positive) value means that the predicted value was greater (less) than the observed. The solid line is the mean, 0.0003, and the dashed line is the median, 0.0216, of the computed differences. The distribution is left-skewed.

TABLE 3 WTWr multiple linear regression model.

Variable	Dependent variable (score)
	HAT Task 1 and 2
WTWs	0.004***
	(0.0003)
WTWr	0.009
	(0.031)
WTWs:WTWr	−0.005***
	(0.001)
Constant	0.685***
	(0.013)
Treatment controls	Yes
Observations	397
R^2	0.402
Adjusted R^2	0.394
RSE	0.115 (df = 391)
F Stat.	52.547*** (df = 5; 391)

*p < 0.1; **p < 0.05; ***p < 0.01.

under and over predictions, respectively) suggests that the two groups had similar mean scores ($t = -0.174$, $df = 36.023$, $p = 0.863$). Similarly, there was not a statistical difference in their WTWs ($t = -1.426$, $df = 23.729$, $p = 0.167$), although those for which the model over-predicted scores, meaning the difference was negative, did tend to have a higher WTWs (about 54) than those for which it under-predicted (about 34).

5.3 WTWr multiple linear regression model

WTWs and WTWt capture a participant's willingness to work on this task individually (s for solo) and in a HAT (t for team). They do not account for an interaction if it were to exist. WTWr serves this purpose. When interacted with WTWs, WTWr illustrates how the relationship between WTWs and a team's success varies at different levels of WTW with the AI relative to a participant's raw WTW as measured by their solo effort. In other words, WTWr moderates the relationship between WTWs and total team score.

The overall WTWr model (Table 3) was statistically significant [$R^2 = 0.402$, $F_{(5,391)} = 52.547$, $p < 0.001$]. The anticipated main effects were present, meaning WTWs significantly predicted total team score ($\beta = 0.004$, $p < 0.001$), WTWr did not ($\beta = 0.009$, $p = 0.769$), and the interaction was significant ($\beta = -0.005$, $p < 0.001$). As with previous models, we reject the null hypothesis that this model does not offer more explanatory value than a controls-only model ($F = 83.299$, $p < 0.001$). Moreover, it outperforms the WTWs-only model, a WTWr-only model, and an additive model (all $p < 0.001$).

Interpretation of the coefficients from linear regression models with interactions is not straightforward. The fact that WTWs is significant in this model merely means that it has an effect in the hypothetical instance when a person's effort with and without the AI was the same. The non-significant result for WTWr means that when WTWs is zero WTWr does not have an effect (this is a soundness check, since that case does not exist in our data). The interaction, however, suggests that we expect WTWr to have an effect at different levels of WTWs. We can get insight into the model's predictions by decomposing the interaction into simple slopes (the slope of WTWs at a set of particular levels of WTWr) and plotting the results.

We chose three representative values of WTWr with which to estimate the slope of WTWs (often called a spotlight analysis): the mean of WTWs plus one standard deviation above and below the mean, as is conventional. The mean WTWr value is 14.2%. The standard deviation values are 40.0% and −12.6%, respectively. The positive mean and standard deviation above the mean values indicate putting in less effort when working with the AI than alone.

The spotlight analysis plot (Figure 3) reveals an increase in the moderating effect of WTWr on WTWs. As WTWr decreases (i.e. a person worked more in the HAT), the predicted relationship between WTWs and total team score also increases. Since none of the plotted confidence intervals contain zero, we can conclude that the slopes are significant for each selected level. The model fit is imperfect, as it predicts impossible scores, i.e., better than 100% for participants with exceptional WTWs (the plot trims these observations). Nevertheless, it still illustrates the moderating effect of willingness to work in a HAT *relative* to willingness to work in general.

FIGURE 3
Spotlight analysis plot of the moderating effect of WTWr on the relationship between WTWs and total team scores. Scores are in percentages, WTWs is truncated at 75 submissions to ignore outliers, and the 95% confidence levels that do not cross zero indicate a significant slope. Note that the expected scores on a multi-peaked landscape are greater than 50% because landscapes are built in constrained spaces. Based on the landscapes generated for this experiment, the expected score of a participant who did no tuning in both trials was about 47%. Submitting just four local searchers would reveal the immediate topology. For the single-peaked landscape, from the valley to the peak was a maximum of 24 1-setting changes to the dials. Thus, a small number of submissions was sufficient for modestly sophisticated participants to achieve scores in the 70-80% range.

5.4 Ordinal logistic regression models

Classifying people is often more useful than predicting the exact outcome of a HAT interaction. The models based on Equation 2, reported in Table 4, do such classification by using WTWs and WTWt to predict the likelihood of a HAT achieving a low, medium, or high score. As outlined above, these classifications are simply the terciles of the observed data.

The overall WTWs ordinal regression model failed the proportional odds assumption, meaning the relationship between each pair of outcome groups (the low, medium, and high HAT score terciles) is not the same. The alternative modeling approach when the proportional odds assumption does not hold is to fit a log-linear model, in this case, a multinomial logistic regression, which we did. Even though the ordinal model failed the proportional odds assumption, the multinomial logistic regression only resulted in a marginally better fit ($AIC = 709.381$ vs. the ordinal logistic model's $AIC = 714.332$). As the proportional odds test is simply one of goodness of fit, and given that the alternative model is not a meaningfully better fit based on the AIC comparison, we opted to stick with the ordinal logistic regression as it is more parsimonious. Based on this model, we again reject the null hypothesis that including WTWs does not contribute to the model with just the treatment controls ($\chi^2 = 152.093, p > 0.001$).

WTWs significantly predicted which category a HAT fell into ($\eta = 0.047, p < 0.001$). The coefficients from ordinal logistic models can be challenging to interpret as they are scaled in terms of logs. Thus, we converted the WTWs coefficient into a proportional odds ratio [$Odds\,Ratio = 1.048$, 95% $CI = (1.038, 1.058)$]. For every one unit increase in a participant's WTWs, the odds of being more successful (medium or high score categories) is multiplied 1.05 times (i.e., increases 5%), all else being equal.

TABLE 4 Ordinal logistic regression models.

Variable	Dependent variable (low, med., or high score class)	
	HAT Task 1 and 2	HAT Task 2
	(1)	(2)
WTWs | WTWt	0.047***	0.083***
	(0.005)	(0.010)
Treatment controls	Yes	Yes
Observations	397	397
Residual deviance	704.332	771.587
AIC	714.332	781.587

*p < 0.05; **p < 0.01; ***p < 0.001.

Another alternative for interpretation of the model is computing predicted probabilities for each outcome category and plotting the values, as in Figure 4. Although there are some minor differences across the treatment conditions, the main effect of WTWs is relatively consistent across the conditions. The probability of the model classifying a HAT as *low* steadily decreases with each additional submission that a participant made during their solo effort (the solid purple line); the inverse is true for the *high* classification (the blue dotted line). The model suggests an increasing probability of being classified as *medium* up to ~50 submissions during the solo effort, at

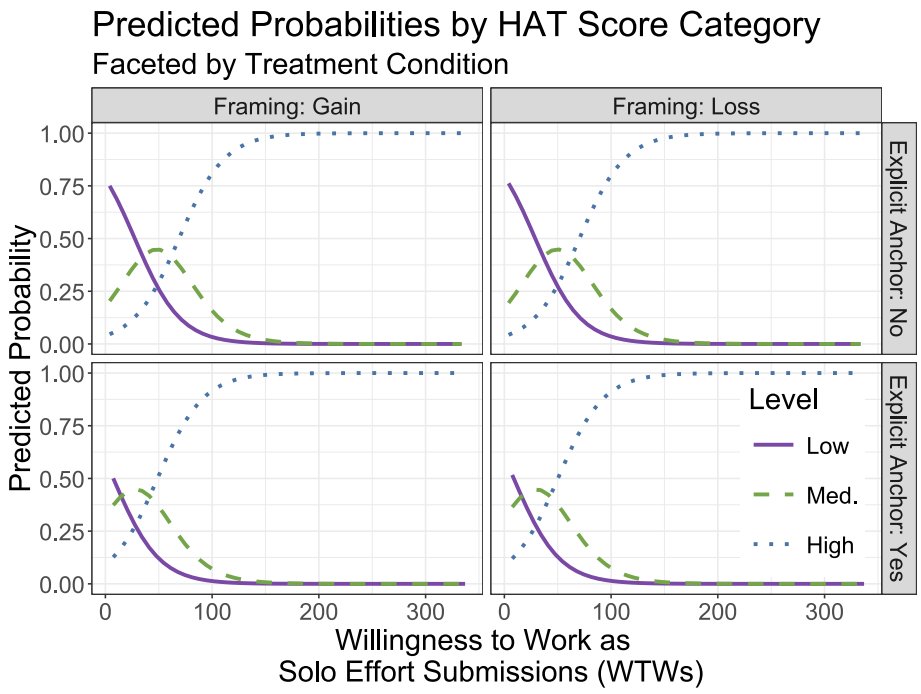

FIGURE 4
A plot of predicted probabilities for categorization into each of the three performance level groups faceted by the experimental treatment conditions. WTWs suggest distinct predicted probability distributions for each of the three categories.

which point the classification probability starts to decline (green dashed line).

The overall WTWt ordinal regression model passed the proportional odds assumption, meaning that the relationships between each pair of outcome groups were not significantly different, i.e., the model was a good fit. This model also rejects the null hypothesis that including WTWt does not contribute to the model with just the treatment controls ($\chi^2 = 84.839, p > 0.001$). WTWt significantly predicted which category a HAT score for the second task belonged to $\eta = 0.083$, $p < 0.001$). We also compute the proportional odds ratio for this coefficient [$Odds\,Ratio = 1.086$, 95% $CI = (1.066, 1.0109)$], which suggests that for every one unit increase in a participant's WTWt, the odds of being more successful (medium or high score categories) during the second task is multiplied 1.09 times (i.e., increases 9%), all else being equal. We also carried out the same predicted probabilities analysis as we did for WTWs. The only notable difference was a leftward shift in the lines, a result of WTWt only including one task instead of two (thus, we would expect roughly half the average value).

6 Discussion

The ability to accurately anticipate the future success of human-agent teams has broad applications, from appropriately marketing the likely value of a new technology to calibrating an AI's model(s) of its own teaming outcomes. Prior work has linked experience to HAT outcomes in a variety of ways. For example, more experienced humans tend to be more skilled and can contribute more skilled effort (Chen and Barnes, 2014), experience in a HAT can improve trust calibration [which is hypothesized to correlate with well-calibrated compliance (Hafizoğlu and Sen, 2018; Mercado et al., 2016; Yang et al., 2017)], and create a sense of group belonging (Savela et al., 2021). Our simple hypothesis was that experience can serve as a proxy measure of a person's willingness to do the HAT task and that a higher willingness to do the task would correlate with better outcomes. We support this hypothesis with analyses of data from a previous experiment.

Our analysis revealed that, as measured by how much effort a study participant put into a task on their own before teaming with an agent, was significantly and positively correlated with their future HAT's success. Moreover, it was equally effective in a categorical model as the linear model, meaning that it has the potential as a tool for binning new HATs according to their potential for success.

Extending this insight, we applied a k-folds cross-validation model. The results suggest robust prediction by the general model and that the outliers may be due to the experimental design. The sample size constrained our ability to verify this effect. It is worth noting that the validation suggested that the model tended to over-predict the scores of participants who displayed a higher willingness to do the task on their own.

We also found that a higher WTW in a HAT after having some experience, as measured by how much effort a study participant put into their first teaming task, was significantly and positively correlated with their HAT's second task success. Again, we fit a categorical model, and it was similarly effective.

Importantly, we used an interaction model to explore the moderating effect of a participant's willingness to work in a HAT on the effect of their raw willingness to do the task. The model suggests that having relatively more willingness to work in a HAT does positively moderate the effect of raw willingness to do the task on a team's success. In other words, a high willingness to do the task solo is most predictive in cases where a person also has a high willingness to do the task in a HAT.

6.1 Interpretation of results

The above results generally support our hypothesis that experience is a reliable predictor of human-agent teaming outcomes because it serves as a proxy for willingness to work. We did not explicitly test, thus debunk, alternative hypotheses. Experience did, however, explain a significant amount of variance in the models that we tested. This is encouraging, given that other hypothesized instruments, e.g., disposition to trust (Gurney et al., 2022b), often come up short. In all likelihood, experience is a useful predictor of human-agent team outcomes for a variety of reasons. Based on our results, we believe a major reason is that it simply serves as a proxy measure of willingness to do a given task. This reason does not undermine the utility of experience in predicting HAT outcomes. However, it does suggest the need for a deeper understanding of the nuances of experience so that models can better map the relationship between past, present, and future HAT interactions.

The willingness to work results reported herein also suggest an opportunity for reevaluating previous work. Many experiments in the HAT space reflect the structure of the data we analyzed: a human was familiarized with a task on their own, gained an intelligent agent teammate, and then worked with the teammate on the task for a given number of events. We see these as excellent opportunities for further vetting our hypothesis and comparing it against prior explanations.

6.2 Related work

6.2.1 Experience and human-agent teams

Experience can impact a HAT in multiple ways (see Table 5 for existing research). Experience with a trustworthy (untrustworthy) agent, for example, was correlated with an increase (decrease) in human trust while teaming with an agent in the Game of Trust (Hafizoğlu and Sen, 2018). Hafizoğlu and Sen argue that this result is due to the humans' emotional states created by their experiences. These lab results are supported by a survey of autonomous driving-capable automobile owners that found that drivers who had experienced unexpected behaviors from their autonomous car reported lower levels of trust in the system (Dikmen and Burns, 2017). Experience can also calibrate expectations: people with experience working with industrial robots were less impressed by a robot's skills than those without (Sarkar et al., 2017). Additionally, experience may help teams avoid uncertainty around roles and task responsibilities, as was found to be the case in robotic surgery teams (Cunningham et al., 2013). A lack of relevant experience with agents, on the other hand, is not necessarily correlated with teaming likelihood, meaning that people with and without experience in a type of HAT are equally likely to join one (Schaefer et al., 2012).

TABLE 5 Representative extant literature.

References	Discipline	HAT	WTW as predictor
Swallow and Woudyalew (1994)	Economics	No	Yes
Sycara and Lewis (2004)	Psychology	Yes	No
Hung et al. (2007)	Policy	No	Yes
Cafaro et al. (2016)	HCI	Yes	No
Gibson et al. (2016)	Economics	No	Yes
You and Robert (2018)	Robotics	Yes	No
Boukaram et al. (2021)	HAI	Yes	No
Paluch et al. (2021)	Management	Yes	No
Verma and Singh (2022)	Management	Yes	No

General, domain-relevant experience can also impact the HAT success. Previous gaming experience, for example, was correlated with how successfully people teamed with robots to capture targets (Chen et al., 2011). Similarly, when trying to identify targets, gaming experience predicted better teaming with a robot that interfaced between a human teammate and less sophisticated robots (Chen and Barnes, 2012). Such findings are likely the result of gaming experience facilitating better visuospatial abilities (Chen and Barnes, 2014).

Humans bring extensive and rich experience into a HAT. Such experience does not, however, guarantee productive outcomes, especially if the teammates are not well-calibrated for their roles (Bhardwaj et al., 2020; Paleja et al., 2021). Within-trial experience is also a useful instrument for understanding HAT dynamics. For example, introducing an agent was correlated with worse performance than human-only in driving (Bhardwaj et al., 2020) and in a collaborative virtual construction task (Paleja et al., 2021) for highly-experienced people. Also, the first compliance decision that participants made in a reconnaissance scenario was a significant predictor of their future compliance behavior (Gurney et al., 2022b). Training AI to have awareness of and the ability to leverage such factors is believed to be essential to getting the most out of HATs (Kamar, 2016).

There are numerous other proposals for how experience can impact a HAT. Examples include experience improving trust (Groom and Nass, 2007; Gutzwiller and Reeder, 2021), experience reducing the probability of humans ignoring an agent after false alarms (tied to a calibration mechanism) (Yang et al., 2017), and the experience creating higher in-group identification with robots (Savela et al., 2021). While all of these are useful, the context-dependent nature of HATs limits their predictive value (Curnin et al., 2015). We argue that WTW is less context-dependent, and as we demonstrated, easy to measure and instrument.

6.2.2 Willingness to work

People exhibit considerable variation in their willingness to engage in different tasks. For example, while some people love working as software engineers, others cannot think of a duller profession. Similarly, while some people look forward to commuting by bike to work, others find it miserable and do whatever they can to avoid it. As a simplification, we call the natural, observable variation between people in their willingness to engage in a given task their WTW. Generally speaking, a person may have an entirely different WTW as a software engineer from their "WTW" as a bike commuter. There are likely many intriguing endogenous and exogenous factors related to a person having a certain WTW. We are primarily interested in its observable manifestation and its correlation with HAT outcome(s). Our basic WTW prediction is quite simple: all else being equal, a person with a higher WTW as a software engineer will measurably do more software engineering. More generally speaking, our definition of WTW positions it as an atomic element of high-order psychological constructs such as motivation (Heckhausen, 1977), grit (perseverance) (Duckworth et al., 2007; Credé et al., 2017), or work ethic (Furnham, 2021).

Outside of our data-driven demonstration, there is evidence that WTW is a correlate of HAT interactions (You and Robert, 2018)—this research, however, posits WTW as an outcome measure and not an independent variable. We hypothesize that a person's basic WTW predicts HAT success, an idea that has been studied in other contexts. For example, in economics, WTW is used as an alternative to willingness to pay because the latter can be much more volatile due to its relationship to wealth (e.g., Swallow and Woudyalew, 1994; Hung et al., 2007; Gibson et al., 2016). These studies use a method known as contingent valuation in which people complete a survey that asks whether they would accept an outcome at different prices or amounts of work. Usually, the outcome is a benefit, but it can also be the avoidance of a loss. Participants indicate what they would be willing to give up to have that benefit—such as how much labor (or money) they would contribute to reducing the presence of a noxious biting fly (Swallow and Woudyalew, 1994). Our review of the literature did not uncover work that identifies a behavioral proxy for such measures, particularly in the case of human-agent teaming. We argue that past effort in the same (or similar) setting is a good candidate behavioral proxy and believe that our demonstration of its effectiveness using data collected to study how human heuristics and biases factor into the performance of new HATs establishes a paradigm for future work.

Other disciplines have tried to capture the idea of a person's willingness to work using psychometric scales. For example, Miller, Woehr, and Hudspeth developed the 65-item Multidimensional Work Ethic Profile (MWEP) to capture how a person relates to the concept of work (Miller et al., 2002). Work ethic, which Miller and company define as a commitment to the value and importance of hard work, shares some traits with WTW. For example, a person high in either is likely going to invest time or mental effort in a work task. However, a person may be willing to do something, particularly in exchange for something like monetary gain, but still have poor work ethic because they do not think the work is important. Despite the conceptual difference the MWEP and similar psychometric inventories have been linked to observed work outcomes that may also be linked to WTW, like generational shifts in workplace values (Cogin, 2012).

It seems plausible that WTW predicts or is correlated with other commonly-studied HAT constructs. For example, considerable effort has gone into understanding the role of trust on HAT outcomes, with a general observed finding being that a human's trust in their AI-enabled counterpart needs to be appropriately calibrated (Chen and Barnes, 2014; Gurney et al., 2022a). Interestingly, people often conflate their own factors, such as prior actions, with an AI's performance—which has been shown to undermine trust (Gurney et al., 2023d). It may be the case that people also conflate their WTW with other features of an AI. For example, a person low in WTW may be disappointed in a poor outcome but not recognize that it was, at least in part, due to their low WTW and instead blame their AI teammate. Our data, unfortunately, do not support such hypothesis testing.

We believe that our work has broad implications across domains of research that consider how humans and AI are integrated, function in teams, and work together to achieve objectives. For example, information systems researchers have documented a general willingness to work *with* AI (e.g., Dennis et al., 2023), but have not considered how a person's basic WTW might factor into such willingness. Disentangling these two constructs could inform new avenues of research related to HAT optimization. Relatedly, emerging research suggests that people respond well to digital human agents, particularly when they look and perform on par with human agents (Seymour et al., 2024). The implication of this insight is that the artificial nature of AI teammates could have a lower impact on HAT outcomes resulting in the relative impact of human teammates' WTW on HATs to increase.

6.3 Limitations

The primary limitation of our work is that it only examines WTW in a single task. Although this task is abstract and sought to study human decision making at a rudimentary level such that findings generalized, it is possible that our results do not extend to every domain. The population sample we relied on may also have impacted our observed outcomes: it is well documented that results from online workers do not always align well with those from lab or field samples (Peer et al., 2022). Lastly, although our sample size was sufficient to demonstrate significant results, a larger, more diverse sample would allow us to model how culture, individual differences, etc. impact the WTW outcome that we document.

6.4 Future work

The data that we relied on for this study are very much a snapshot in time. They do not, for example, allow us to control for additional teaming experiences that people may have, both with other humans and autonomous agents. Similarly, even though the tasks were distinct, unique experiences, they happened in a relatively brief period. A longer time horizon with more interactions might lead to different, more nuanced effects. The data are also from a single, abstract task. Controlling for richer

teaming experiences, increasing the time horizon and number of interactions, and looking at contextualized settings, we believe, are all necessary. Importantly, this would facilitate comparing different levels of experience in a fully between design. Finally, data that support both willingness to work and alternative predictors, for example, disposition to trust, would greatly help identify the role played by WTW relative to other human factors.

Additionally, future research should examine the relationship between WTW and work quality. Our current measure captures effort investment through submission counts but does not assess whether higher WTW correlates with work effectiveness per submission. While our results show higher WTW predicts better team outcomes overall, determining whether this stems from work quantity alone or also reflects quality would require larger samples to build models that can classify high and low quality work while controlling for random factors like luck in the landscape search task. Such analysis could reveal whether WTW serves primarily as a measure of willingness to invest effort or if it also captures aspects of individual differences in work effectiveness, helping to better understand its role in predicting human-agent team success. For example, it may be the case that more effective workers have a higher WTW due to their ability to capture rewards. Alternatively, workers with higher WTW may become more effective, further reinforcing their WTW.

Further research should also explore how the WTW paradigm translates to more dynamic, interdependent tasks. Our study used a relatively structured dial-tuning task with clear turn-taking between human and agent. However, many real-world human-agent teams operate in environments requiring continuous coordination, simultaneous actions, and complex interdependencies between teammates. For example, search and rescue scenarios (Pynadath et al., 2023), collaborative manufacturing (Seeber et al., 2020), or real-time strategic planning (Narne et al., 2024) all involve more fluid interaction patterns. Understanding how WTW manifests and predicts team success in such contexts would be valuable. This could include developing new measures of WTW suitable for dynamic tasks, examining how task interdependence affects the relationship between WTW and team outcomes, and investigating whether WTW remains a stable predictor across different types of human-agent coordination patterns.

7 Conclusion

We explored one way a person's experience in a task domain can correlate with the dynamics and impact the success of the human-agent teams they enter. Although previous research posits many different mechanisms by which experience can hold such sway over HATs, it largely ignores the simple insight that we present: measures of experience may simply be serving as proxies for willingness to do the task. Although this hypothesis cannot explain every instance of correlation between experience and human-agent teaming outcomes, we argue that it can account for many. Moreover, it is simple to assess, is robust enough for prediction (thus useful to agents in modeling their human counterparts), and suggests a need for reevaluation of other empirical results related to experience and teaming success.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by University of Southern California Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

NG: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. DP: Writing – review & editing. JM: Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The project or effort depicted was or is sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2025.1405436/full#supplementary-material

References

Bhardwaj, A., Ghasemi, A. H., Zheng, Y., Febbo, H., Jayakumar, P., Ersal, T., et al. (2020). Who's the boss? Arbitrating control authority between a human driver and automation system. *Transp. Res. F Traffic Psychol. Behav*. 68, 144–160. doi: 10.1016/j.trf.2019.12.005

Boukaram, H-. A., Ziadee, M., and Sakr, M. F. (2021). "Mitigating the effects of delayed virtual agent response time using conversational fillers," in *Proceedings of the 9th International Conference on Human-Agent Interaction* (New York, NY: ACM), 130–138. doi: 10.1145/3472307.3484181

Cafaro, A., Vilhjálmsson, H. H., and Bickmore, T. (2016). First impressions in human-agent virtual encounters. *ACM Trans. Comput.-Hum. Interact*. 23, 1–40. doi: 10.1145/2940325

Chapman, G. B., and Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organ. Behav. Hum. Decis. Process*. 79, 115–153. doi: 10.1006/obhd.1999.2841

Chen, J. Y., and Barnes, M. J. (2012). Supervisory control of multiple robots: effects of imperfect automation and individual differences. *Hum. Factors* 54, 157–174. doi: 10.1177/0018720811435843

Chen, J. Y., and Barnes, M. J. (2014). Human-agent teaming for multirobot control: a review of human factors issues. *IEEE Trans. Hum-Mach. Syst*. 44, 13–29. doi: 10.1109/THMS.2013.2293535

Chen, J. Y., Barnes, M. J., Quinn, S. A., and Plew, W. (2011). Effectiveness of roboleader for dynamic re-tasking in an urban environment. *Proc. Hum. Factors Ergon. Soc. Annu. Meet*. 55, 1501–1505. doi: 10.1177/1071181311551312

Cogin, J. (2012). Are generational differences in work values fact or fiction? Multi-country evidence and implications. *Int. J. Hum. Resour. Manag*. 23, 2268–2294. doi: 10.1080/09585192.2011.610967

Credé, M., Tynan, M. C., and Harms, P. D. (2017). Much ado about grit: a meta-analytic synthesis of the grit literature. *J. Pers. Soc. Psychol*. 113:492. doi: 10.1037/pspp0000102

Cunningham, S., Chellali, A., Jaffre, I., Classe, J., and Cao, C. G. (2013). Effects of experience and workplace culture in human-robot team interaction in robotic surgery: a case study. *Int. J. Soc. Robot*. 5, 75–88. doi: 10.1007/s12369-012-0170-y

Curnin, S., Owen, C., Paton, D., Trist, C., and Parsons, D. (2015). Role clarity, swift trust and multi-agency coordination. *J. Contingencies Crisis Manag*. 23, 29–35. doi: 10.1111/1468-5973.12072

Demir, M., Cooke, N. J., and Amazeen, P. G. (2018). A conceptual model of team dynamical behaviors and performance in human-autonomy teaming. *Cogn. Syst. Res*. 52, 497–507. doi: 10.1016/j.cogsys.2018.07.029

Dennis, A. R., Lakhiwal, A., and Sachdeva, A. (2023). AI agents as team members: EFFECTS on satisfaction, conflict, trustworthiness, and willingness to work with. *J. Manag. Inform. Syst*. 40, 307–337. doi: 10.1080/07421222.2023.2196773

Dikmen, M., and Burns, C. (2017). "Trust in autonomous vehicles: the case of tesla autopilot and summon," in *2017 IEEE International conference on systems, man, and cybernetics (SMC)* (Banff, AB: IEEE), 1093–1098. doi: 10.1109/SMC.2017.8122757

Duckworth, A. L., Peterson, C., Matthews, M. D., and Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *J. Pers. Soc. Psychol*. 92:1087. doi: 10.1037/0022-3514.92.6.1087

Epley, N., and Gilovich, T. (2006). The anchoring-and-adjustment heuristic: why the adjustments are insufficient. *Psychol. Sci*. 17, 311–318. doi: 10.1111/j.1467-9280.2006.01704.x

Furnham, A. (2021). *The Protestant Work Ethic: The Psychology of Work Related Beliefs and Behaviours*. London: Routledge doi: 10.4324/9781003209126

Gibson, J., Rigby, D., Polya, D., and Russell, N. (2016). Discrete choice experiments in developing countries: willingness to pay versus willingness to work. *Environ. Resour. Econ*. 65, 697–721. doi: 10.1007/s10640-015-9919-8

Groom, V., and Nass, C. (2007). Can robots be teammates?: benchmarks in human-robot teams. *Interact. Stud*. 8, 483–500. doi: 10.1075/is.8.3.10gro

Gurney, N., King, T., and Miller, J. H. (2022a). "An experimental method for studying complex choices," in *International Conference on Human-Computer Interaction* (Cham: Springer), 39–45. doi: 10.1007/978-3-031-19679-9_6

Gurney, N., Miller, J., and Pynadath, D. (2023a). *The role of heuristics and biases in complex choices*. Preprint. doi: 10.21203/rs.3.rs-2472194/v1

Gurney, N., Miller, J. H., and Pynadath, D. V. (2023b). The role of heuristics and biases during complex choices with an ai teammate. *Proc. AAAI Conf. Artif. Intell*. 37, 5993–6001. doi: 10.1609/aaai.v37i5.25741

Gurney, N., Pynadath, D. V., and Wang, N. (2022b). "Measuring and predicting human trust in recommendations from an ai teammate," in *International Conference on Human-Computer Interaction* (Cham: Springer), 22–34. doi: 10.1007/978-3-031-05643-7_2

Gurney, N., Pynadath, D. V., and Wang, N. (2023c). "Comparing psychometric and behavioral predictors of compliance during human-ai interactions," in *Persuasive Technology: 18th International Conference, PERSUASIVE 2023, Eindhoven, The Netherlands, April 19–21, 2023, Proceedings* (Cham: Springer), 175–197. doi: 10.1007/978-3-031-30933-5_12

Gurney, N., Pynadath, D. V., and Wang, N. (2023d). My actions speak louder than your words: when user behavior predicts their beliefs about agents' attributes. In International Conference on Human-Computer Interaction (Springer), 232–248. doi: 10.1007/978-3-031-35894-4_17

Gutzwiller, R. S., and Reeder, J. (2021). Dancing with algorithms: interaction creates greater preference and trust in machine-learned behavior. *Hum. Factors* 63, 854–867. doi: 10.1177/0018720820903893

Hafizoğlu, F. M., and Sen, S. (2018). "The effects of past experience on trust in repeated human-agent teamwork," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm), 514–522

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

Heckhausen, H. (1977). Achievement motivation and its constructs: a cognitive model. *Motiv. Emot*. 1, 283–329. doi: 10.1007/BF00992538

Hoff, K. A., and Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Hum. Factors* 57, 407–434. doi: 10.1177/0018720814547570

Huang, H-. Y., and Bashir, M. (2017). "Personal influences on dynamic trust formation in human-agent interaction," in *Proceedings of the 5th International Conference on Human Agent Interaction* (New York, NY: ACM), 233–243. doi: 10.1145/3125739.3125749

Hung, L. T., Loomis, J. B., and Thinh, V. T. (2007). Comparing money and labour payment in contingent valuation: the case of forest fire prevention in Vietnamese context. *J. Int. Dev*. 19, 173–185. doi: 10.1002/jid.1294

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

Kamar, E. (2016). "Directions in hybrid intelligence: complementing ai systems with human intelligence," in *IJCAI* (New York, NY), 4070–4073

Kuhn, M. (2008). Building predictive models in r using the caret package. *J. Stat. Softw*. 28, 1–26. doi: 10.18637/jss.v028.i05

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Factors* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Lewis, M., Sycara, K., and Walker, P. (2018). "The role of trust in human-robot interaction," in *Foundations of Trusted Autonomy*, eds. H. Abbass, J. Scholz, and D. Reid (Cham: Springer), 135–159. doi: 10.1007/978-3-319-64816-3_8

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

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., Procci, K., et al. (2016). Intelligent agent transparency in human-agent teaming for multi-uxv management. *Hum. Factors* 58, 401–415. doi: 10.1177/0018720815621206

Miller, M. J., Woehr, D. J., and Hudspeth, N. (2002). The meaning and measurement of work ethic: construction and initial validation of a multidimensional inventory. *J. Vocat. Behav*. 60, 451–489. doi: 10.1006/jvbe.2001.1838

Narne, S., Adedoja, T., Mohan, M., and Ayyalasomayajula, T. (2024). AI-driven decision support systems in management: enhancing strategic planning and execution. *Int. J. Recent Innov. Trends Comput. Commun*. 12, 268–276.

O'Neill, T., McNeese, N., Barron, A., and Schelble, B. (2022). Human-autonomy teaming: a review and analysis of the empirical literature. *Hum. Factors* 64, 904–938. doi: 10.1177/0018720820960865

Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., and Gombolay, M. (2021). The utility of explainable AI in *ad hoc* human-machine teaming. *Adv. Neural Inform. Process. Syst*. 34, 610–623.

Paluch, S., Tuzovic, S., Holz, H. F., Kies, A., and Jörling, M. (2021). "My colleague is a robot"-exploring frontline employees' willingness to work with collaborative service robots. *J. Serv. Manag*. 33, 363–388. doi: 10.1108/JOSM-11-2020-0406

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., and Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods* 54, 1643–1662. doi: 10.3758/s13428-021-01694-3

Pynadath, D., Gurney, N., Kenny, S., Kumar, R., Marsella, S., Matuszak, H., et al. (2023). "Effectiveness of teamwork-level interventions through decision-theoretic reasoning in a minecraft search-and-rescue task," in *Proceedings of the 2023 International Conference On Autonomous Agents And Multiagent Systems* (London), 2334–2336.

Sarkar, S., Araiza-Illan, D., and Eder, K. (2017). Effects of faults, experience, and personality on trust in a robot co-worker. *arXiv* [Preprint]. arXiv:1703.02335. doi: 10.48550/arXiv.1703.02335

Savela, N., Kaakinen, M., Ellonen, N., and Oksanen, A. (2021). Sharing a work team with robots: the negative effect of robot co-workers on in-group identification with the work team. *Comput. In Hum. Behav*. 115:106585. doi: 10.1016/j.chb.2020.106585

Schaefer, K. E., Sanders, T. L., Yordon, R. E., Billings, D. R., and Hancock, P. A. (2012). "Classification of robot form: factors predicting perceived trustworthiness," in *Proceedings of the human factors and ergonomics society annual meeting. Vol. 56* (Los Angeles, CA: SAGE Publications Sage CA), 1548–1552. doi: 10.1177/1071181312561308

Seeber, I., Bittner, E., Briggs, R. O., De Vreede, T., De Vreede, G. J., Elkins, A.B., et al. (2020). Machines as teammates: a research agenda on AI in team collaboration. *Inform. Manag*. 57:103174. doi: 10.1016/j.im.2019.103174

Seo, M., Barrett, L., and Bartunek, J. (2004). The role of affective experience in work motivation. *Acad. Manag. Rev*. 29, 423–439. doi: 10.2307/20159052

Seymour, M., Yuan, L., Riemer, K., and Dennis, A. R. (2024). Less artificial, more intelligent: understanding affinity, trustworthiness, and preference for digital humans. *Inf. Syst. Res*. doi: 10.1287/isre.2022.0203. [Epub ahead of print].

Swallow, B. M., and Woudyalew, M. (1994). Evaluating willingness to contribute to a local public good: application of contingent valuation to tsetse control in Ethiopia. *Ecol. Econ*. 11, 153–161. doi: 10.1016/0921-8009(94)90025-6

Sycara, K., and Lewis, M. (2004). "Integrating intelligent agents into human teams," in *Team Cognition: Understanding the Factors that Drive Process and Performance*, eds. E. Salas, and S. M. Fiore (Washington, DC: American Psychological Association), 203–231. doi: 10.1037/10690-010

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases: biases in judgments reveal some heuristics of thinking under uncertainty. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124

Tversky, A., and Kahneman, D. (1985). "The framing of decisions and the psychology of choice," in *Behavioral decision making* (Cham: Springer), 25–41. doi: 10.1007/978-1-4613-2391-4_2

Verma, S., and Singh, V. (2022). The employees intention to work in artificial intelligence-based hybrid environments. *IEEE Trans. Eng. Manag*. 71, 3266–3277. doi: 10.1109/TEM.2022.3193664

Yang, X. J., Unhelkar, V. V., Li, K., and Shah, J. A. (2017). "Evaluating effects of user experience and system transparency on trust in automation," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (New York, NY: IEEE), 408–416. doi: 10.1145/2909824.3020230

You, S., and Robert, L. P. (2018). "Human-robot similarity and willingness to work with a robotic co-worker," in *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (New York, NY: IEEE), 251–260. doi: 10.1145/3171221.3171281