Perceived Teammate Identity and Cooperative Behavior in Hybrid Human–Artificial Agent Teams [†]

Amrote Seyoum Getu[a,**], Radosveta Ivanova-Stenzel[a,b,**], Michel Tolksdorf[a,b,*], Eva Wiese[a,**]

[a] *Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany*
[b] *Berlin School of Economics*

Abstract

We study how perceived teammate identity shapes effort allocation in hybrid human–artificial agent teams in collaborative environments. In a real-effort online experiment and under different incentive schemes, fixed payment and team competition, participants work in a team where each team member is responsible for their own segment but can observe and intervene in teammates' segments. Teammates are algorithmically controlled agents, framed as either artificial or human-like, with their actual behavior held constant across treatments. Participants exert approximately one-third less compensatory effort when teammates are presented with human-like cues than when they are transparently labeled as artificial, and this difference is robust across both incentive schemes. The pattern is consistent with a model in which perceived human identity reduces the non-monetary benefit of helping. The result highlights that in hybrid teams, the labeling and presentation of artificial collaborators have economically meaningful consequences for cooperative effort and the organization of team production.

Keywords: Human–AI Collaboration, Online Experiment, Team Incentives, Teamwork, Tournaments
JEL classifications: C91, D29, D83, D89

1. Introduction

Artificial intelligence (AI) technologies are increasingly integrated into production and decision environments that were once the sole domain of human labor (Cioffi et al., 2020). This development marks a shift toward hybrid teams, where humans and artificial agents jointly determine output and efficiency. While AI systems can enhance accuracy and reduce processing costs, they also alter the behavioral allocation of effort within teams. Understanding how humans adjust their actions and coordination strategies in the presence of artificial partners is essential for predicting the productivity and welfare effects of hybrid work. This challenge is amplified in digital environments, where people commonly interact without knowing whether the partner on the other side is a real human or an AI-driven agent.

Assuming rationality and self-interest, neither the team composition nor the nature of its members should influence how effort is allocated. In practice, however, the decision to intervene is shaped by behavioral mechanisms such as social norms, perceptions of reciprocity, and felt obligations toward teammates, which

[*] Corresponding author: michel.tolksdorf@tu-berlin.de
[**] a.getu@tu-berlin.de, ivanova-stenzel@tu-berlin.de, eva.wiese@tu-berlin.de

[†] Financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 "Rationality and Competition" (project number 280092119) is gratefully acknowledged. Data will be made available upon request. Declarations of interest: none.

are sensitive to who those teammates are believed to be. This is especially apparent in collaborative settings where team members are each responsible for their own area but can observe when a teammate falls short and must decide whether to incur the cost of stepping in to help. Such settings arise in a wide range of environments, including healthcare units, manufacturing quality-control systems, air-traffic control centers, cybersecurity monitoring teams, warehouse operations, and customer support organizations.

When teams consist entirely of humans, members often reallocate effort by providing additional support to underperforming teammates, thereby stabilizing collective output (Kandel and Lazear, 1992; Che and Yoo, 2001; Hamilton et al., 2003). In hybrid human–artificial agent teams, however, it is not obvious whether this pattern is amplified or attenuated, as the introduction of artificial teammates activates countervailing behavioral forces. On the one hand, humans may view artificial agents as tools rather than teammates and thus feel less social obligation to step in when an artificial partner underperforms (Walliser et al., 2019). On the other hand, when a teammate is perceived as artificial, individuals may be less willing to offload responsibility for poor outcomes onto that agent than they would be with a human colleague (Gazit et al., 2023), creating a stronger felt obligation to monitor and correct the artificial teammate's behavior. Social norms governing workplace interaction may further suppress intervention when teammates are perceived as human: stepping into a colleague's area of responsibility risks being seen as intrusive or disrespectful, a concern that does not arise when the teammate is perceived as an artificial agent lacking the kind of mind that would register such social violations (Wiese et al., 2022). Because individuals are often more sensitive to algorithmic errors than to human mistakes, a phenomenon known as algorithm aversion (Dietvorst et al., 2015), they may intervene more readily when AI teammates make visible errors. Whether these countervailing forces produce more or less compensatory effort in hybrid teams, and whether the answer depends on how artificial teammates are labeled and presented, is an empirical question that the existing literature has not addressed.

In this paper, we experimentally identify the causal effects of perceived teammate identity on cooperative effort in hybrid teams. In a between-subject design, we vary whether artificial teammates are transparently labeled as such or presented with human-like cues that leave their nature ambiguous. We focus on the behavioral consequences of perceived teammate identity, particularly how it shapes the allocation of cooperative effort within the team.

Our objective is not to study the technological capabilities of advanced AI systems per se, but to isolate the behavioral consequences of collaborating with teammates perceived as artificial rather than human. Thus, the artificial teammates in our experiment are not representing fully autonomous generative AI systems. Rather, they are algorithmically controlled teammates with fixed behavioral rules whose identity is experimentally manipulated through framing. This design choice provides a key advantage for identification. Because the underlying behavior of the teammates is held constant across treatments, any observed differences in intervention behavior or effort allocation can be causally attributed to perceived teammate identity rather than differences in actual teammate quality, adaptiveness, or strategic sophistication.

In addition, we investigate whether the effects of perceived teammate identity persist under alternative incentive structures. Beyond a condition in which participants are compensated using a fixed payment scheme, we also study competition between teams, where the highest-performing teams receive a bonus. Previous research on team production and intergroup competition finds that competitive incentives raise effort relative to fixed payment schemes. The effect of intergroup competition has been studied, for

example, in the context of the step-level public goods game with and without communication (Bornstein, 1992), the prisoner's dilemma game (Bornstein and Ben-Yossef, 1994), the minimum-effort game (Bornstein et al., 2002), price competition (Bornstein and Gneezy, 2002), and non-routine team tasks (Englmaier et al., 2024). However, none of the papers on intergroup competition address how such incentives influence behavior when teams consist partly of artificial rather than exclusively human members.

Our results show that participants exert significantly less compensatory effort when teammates are presented with human-like cues that leave their nature ambiguous, and this effect is robust across both incentive schemes. Under fixed payment, intervention is significantly lower when teammates are presented with anthropomorphic cues, leaving participants uncertain about whether they are interacting with human or artificial agents. This difference persists under intergroup competition, indicating that the belief-induced shift in cooperative effort reflects a stable behavioral response to perceived teammate identity rather than an artifact of a particular incentive environment. The pattern is also consistent with a model in which the belief that teammates are human reduces the non-monetary benefit of helping.

A recent but rapidly growing literature examines how artificial agents function within human teams and how their presence reshapes coordination and performance. Prior research shows that AI teammates can improve information integration but also introduce new coordination frictions (Zercher et al., 2025), that effective human–AI collaboration requires shared situational awareness and well-defined roles (Hagemann et al., 2023), and that individuals engage in high levels of cooperation with AI systems such as large language models (Niszczota et al., 2025). Recent studies also highlight how the introduction of AI alters team structure, determining which human roles are most likely to be replaced (Cheng et al., 2025), and how hybrid teams can outperform human-only teams under certain conditions when AI agents complement human capabilities (Dell'Acqua et al., 2025). While this literature provides important insights into AI adoption, team composition, coordination, and performance, it does not examine how the perceived identity of artificial teammates shapes cooperative effort within hybrid teams, a gap our experiment fills by providing causal evidence on how perceived teammate identity, and the incentive structure in which it operates, determines the willingness to incur the cost of helping when a teammate underperforms.

Our study is also related to recent work on artificial agents in public-good and threshold public-good environments (Qiao et al., 2026). Like this literature, we examine whether the presence or framing of artificial teammates changes individuals' willingness to incur private costs for collective outcomes. Our experiment complements threshold public-good studies by showing how AI framing affects costly intervention behavior in a setting where cooperation takes the form of reallocating attention and effort toward teammates' areas of responsibility.

2. Experimental Design

In an online experiment, we implemented two treatments that differed only in how the two teammates were framed: in the *AI-labeled* treatment, teammates were presented as artificial agents using distinct icons and system-generated labels prefixed with "[Bot]"; in the *Human-framed* treatment, the same agents were presented using human avatars and names that were also available to participants. In both treatments, the teammates were programmed agents. The manipulation involved only whether these agents were presented with explicit bot labels or with human-like names and avatars. The AI-labeled treatment constitutes a certain-identity baseline of our design, as participants were informed that their teammates were bots. The

Human-framed treatment captures a situation increasingly common in real-world digital environments, where interaction partners are identified only by cues such as avatars and names that do not provide full certainty about whether the counterpart is human or artificially controlled.

The framing allows us to isolate the behavioral effects of perceived teammate identity. Because teammate behavior is held constant across treatments, any differences in participant behavior can be interpreted as a pure belief effect rather than a response to differences in actual teammate performance. We focus on the behavioral consequences that arise from this belief distortion, particularly on how it influences cooperative behavior in hybrid human–artificial agent teams. At the end of the experiment, we asked participants about the perceived nature of their teammates on a 7-point Likert-scale (1 – very unlikely AI, 4 – uncertain, 7 – very likely AI).

A potential concern with any framing manipulation is that participants may infer different experimenter expectations from the labels assigned to teammates. Several design features mitigate this concern. The part of the instruction describing the possibility of helping teammates was identical in both treatments. Moreover, intervention was costly, since selecting targets in a teammate's segment yielded fewer points than selecting targets in one's own segment and mistakes in a teammate's segment were penalized more heavily. Finally, teammate behavior and incentives were identical across treatments.

We adopt a real-effort task in which participants exert measurable effort under limited time within a team of three members (the participant and two programmed agents), making team performance directly dependent on how participants allocate their actions. This design ensures that all observed behavioral adjustments, such as focusing on one's own performance versus intervening and helping teammates, are costly and consequential, thereby capturing economically meaningful effort allocation.

In the experiment, participants performed a collaborative visual search task, identifying visual targets within a shared interface.[1] The task was presented as an online game, in which a circular field was divided into three equal segments, each representing one team member's search area. Across the circle, 21 items appeared (3 targets and 18 distractors), varying in color (blue/green), shape (T/L), and orientation. Each participant controlled their own segment while monitoring the segments of their two teammates (see Figure 2 in the Appendix for a screenshot of the task.)

The goal was to find as many targets as possible to maximize the team's collective score. The scoring system introduced a trade-off between focusing on one's own area and intervening in teammates' areas, reflecting the coordination dilemmas common in team production. Points were awarded (deducted) based on where a target (distractor) was selected: selecting a target (distractor) in one's own segment earned +10 (-5) points, while selecting a target (distractor) in a teammate's segment earned +5 (-10) points.

Consequently, the team score is the sum of the participant's score and the agents' scores. The participant's score can be decomposed into the individual score (points earned in one's own segment) and the intervention score (points earned in both agents' segments):

$$\text{Team Score} = \underbrace{\text{Individual Score} + \text{Intervention Score}}_{\text{Participant's Score}} + \text{Agents' Scores}. \tag{1}$$

[1] The task was adapted from the compound visual search paradigm by Wolfe (1998) and programmed in Unity (v2022.3.8f1).

The agents' actions were automated and based on the same underlying scoring system. To ensure that participants' intervention decisions could be cleanly measured and attributed, the agents were programmed such that (i) they never selected targets or distractors in the participant's segment or the other agent's segments; (ii) they selected only targets, never distractors, in their own segment; and (iii) they missed a percentage of the targets in their own segment.[2]

The experiment consisted of four blocks of 50 trials, with each trial lasting 3 seconds and separated by 2-second breaks.[3] Participants first completed a practice round and then proceeded through all four experimental blocks. The maximum team score that could be achieved in a block was 1400, the maximum individual score was 500, and the maximum (optimal) intervention score was 500 (100). Without intervention, the agents earn 800 points and miss 20 targets per block. Therefore, the player can optimally earn $500 + 100$ points by selecting only their own targets and the targets the agents miss (yielding a maximum team score of 1400).

We implemented two payment conditions to test the robustness of the results across different incentive schemes. In the Fixed-payment condition, participants received a flat payment for completing the task, independent of their team's performance. In the Team-competition condition, a tournament incentive scheme was introduced to induce competitive motivation. Under this scheme, an additional bonus was awarded to the top 20% of participants, ranked according to the total team scores. This incentive structure maintained within-team cooperation while introducing inter-team competition through relative performance incentives.

A total of 289 participants, 144 (145) in Fixed-payment (Team-competition), were recruited through Prolific, an online platform for behavioral research. The participants received £6 for completing the task. In the team-competition condition, those in the top 20% earned an additional £1 bonus. The target sample size was determined based on available funding and recruitment constraints. Within each payment condition, participants were randomly assigned to one of the two treatments: collaboration with two artificial teammates or two human-framed teammates. All participants reported fluent English proficiency. The share of female participants was 39.84% in the Fixed-payment condition and 49.23% in the Team-competition condition. Ages range from 20 to 80 (Mean = 37.52, SD = 11.42) in the Fixed-payment condition and from 21 to 75 (Mean = 39.11, SD = 12.18) in the Team-competition condition. The study received ethical approval from the Ethics Committee of Faculty VII, Economics and Management, at Technische Universität Berlin (project number 20240108). Section 5.2 in the Appendix presents the on-screen instructions.

3. Results

We begin with a brief overview of the experimental outcomes. The team score, defined in equation (1), has three components: the participant's individual score, the participant's intervention score, and the agents'

[2]One of the agents missed 10% and the other 30% of its targets. The programmed variation in miss rates between the two teammates simulates different levels of performance.

[3]The duration of a trial was determined after extensive piloting. It is long enough for participants to identify and select a target in their own segment and to observe what is happening in their teammates' segments, but short enough to create meaningful time pressure and to preserve the core trade-off between focusing on one's own segment and allocating attention to helping teammates.

scores. Agents follow fixed programmed behavior that is identical across treatments. Any treatment effect on team performance therefore originates from the participant's decisions: how much effort to allocate to their own segment (individual score) and how much compensatory effort to direct toward teammates' segments (intervention score). Although intervention may affect realized agent scores when a participant selects a target that an agent would otherwise have selected, participants cannot influence the agents' underlying behavior. We therefore focus our analysis on the participant's individual and intervention scores, which fully capture the behavioral response to perceived teammate identity. Summary statistics for all score variables are presented in Table 1.

		Human-framed treatment					AI-labeled treatment				
Condition	Variable	Mean	SD	Min	Max	Obs.	Mean	SD	Min	Max	Obs.
Fixed-payment	Individual score	412.43	63.87	60	500	288	417.16	57.64	130	500	252
	Intervention score	32.36	39.27	-50	170	288	48.29	49.72	0	205	252
	Agents' scores	769.13	44.79	550	800	288	743.65	82.18	300	800	252
	Team score	1213.92	66.51	800	1330	288	1209.11	88.52	615	1335	252
Team-competition	Individual score	415.45	68.25	30	500	256	423.99	63.00	55	500	272
	Intervention score	35.35	47.56	-10	205	256	52.74	44.01	-60	190	272
	Agents' scores	764.38	59.38	500	800	256	750.00	52.96	550	800	272
	Team score	1215.18	70.02	830	1340	256	1226.73	67.54	855	1335	272

Table 1: Summary Statistics of Team, Individual, Intervention, and Agent Score

3.1. Behavioral Results

Figure 1 presents the mean individual score and the mean intervention score, separated by treatment and incentive condition. There is a persistent and substantial gap in intervention scores between the AI-labeled and Human-framed treatments: participants working with explicitly labeled bot teammates intervene considerably more. Under fixed payment, the mean intervention score is 32.36 in the Human-framed treatment and 48.29 in the AI-labeled treatment (see Table 1). Under competition, the corresponding means are 35.35 and 52.74, respectively. In both incentive conditions, the intervention score in the Human-framed treatment is approximately one third lower than in the AI-labeled baseline. This effect is economically substantial, remarkably stable across all four blocks, and already present from the first block under both incentive schemes. In block 1, mean intervention scores are 43.10 in the AI-labeled treatment and 22.25 in the Human-framed treatment under fixed payment, a difference that persists without systematic trend across subsequent blocks. A corresponding pattern holds under team competition (AI-labeled treatment: 51.47; Human-framed treatment: 30.31). The immediate emergence of this gap in block 1 is particularly notable. To assess whether the treatment difference is already present before any learning or adjustment can take place, we estimate the treatment effect separately for block 1 and find a statistically significant difference (see Table 4 in the Appendix), confirming that the behavioral response reflects a direct reaction to perceived teammate identity rather than a learned adjustment. Individual scores do not differ significantly between treatments under either incentive condition. We applied exclusion criteria at the participant level using block-level performance. Specifically, participants whose own-segment target selection rate or own-segment distractor selection rate was more than two standard deviations from the

corresponding block-specific sample mean were removed from the analysis. We additionally excluded one participant who reported being colorblind. This resulted in a final sample of 134 (132) participants in the Fixed-payment (Team-competition) condition. The results are robust to using other exclusion criteria such as a three standard deviation threshold or to excluding participants based on the overall rather than block-specific performance.

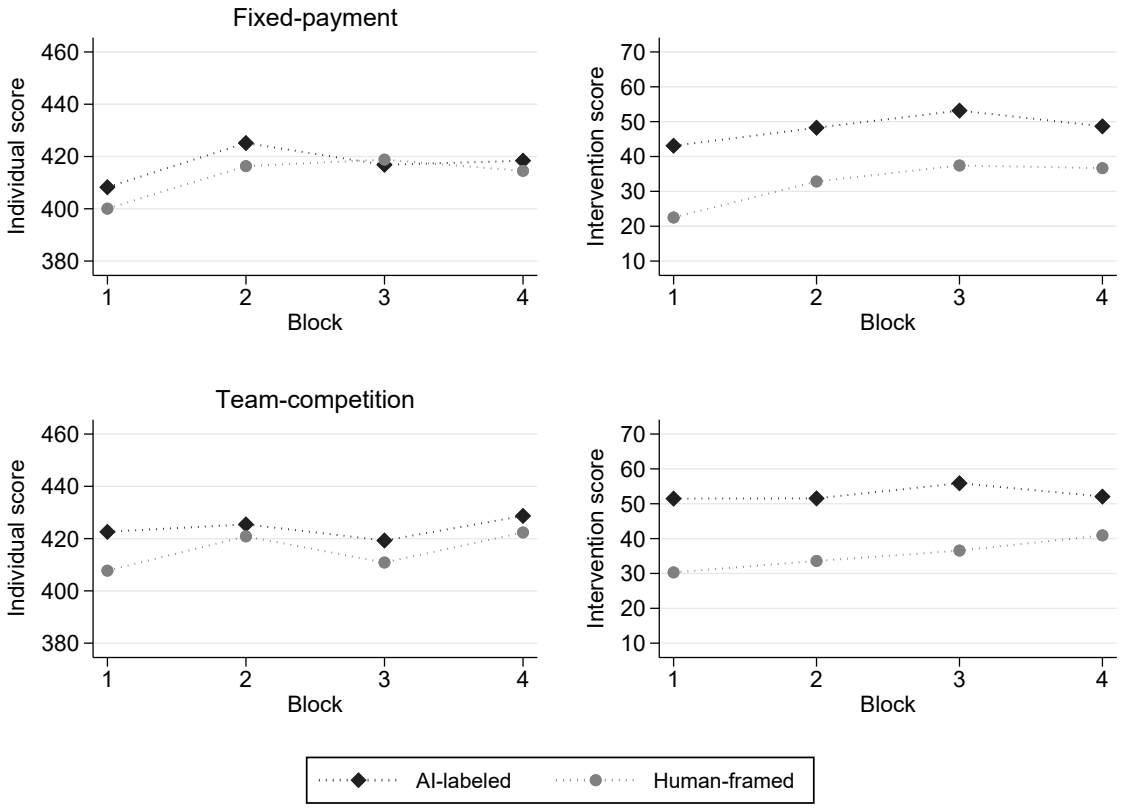

Figure 1: Mean Participants' score (individual score and intervention score) over blocks by treatment.

Specifications (1) and (3) in Table 2 present random-effects regressions of the individual score and intervention score on the treatment indicator (*Human-framed* = 1), controlling for gender, age, and block fixed effects. Under fixed payment (Panel A), being assigned to the Human-framed treatment reduces the intervention score by 15.90 points ($p = .022$), while the individual score is not significantly affected ($p = .603$). Under team competition (Panel B), the reduced intervention score in the Human-framed treatment is 19.65 points ($p = .005$). The intervention result is thus robust across both incentive schemes, confirming that the willingness to exert compensatory effort is significantly lower when teammates are presented with human avatars and names than when they are transparently labeled as bots. Age is negatively associated with the outcomes in most specifications (exact estimates are reported in Table 2), consistent with the gamified nature of the task. Under team competition, female participants achieve significantly lower individual scores ($p = .002$), consistent with a well-established literature documenting that women

perform less well under competitive incentive schemes (Niederle and Vesterlund, 2007; Gneezy et al., 2003; Croson and Gneezy, 2009). Gender is included as a control variable throughout. We do not find evidence that it interacts systematically with the belief channel.

	Individual Score		Intervention Score	
	(1) RE	(2) 2SLS	(3) RE	(4) 2SLS
Panel A: Fixed-payment				
Human-framed	-2.45		-15.68*	
	[8.429]		[6.962]	
Human-likeness		-12.09		-77.26*
		[41.318]		[34.746]
Gender (female)	-9.71	-8.40	-7.08	1.26
	[8.618]	[8.131]	[6.942]	[7.666]
Age	-1.10**	-1.09**	-0.69*	-0.63*
	[0.382]	[0.389]	[0.267]	[0.307]
Constant	451.56***	452.73***	69.08***	76.56***
	[16.800]	[18.135]	[13.034]	[15.272]
Block-fixed effects	✓	✓	✓	✓
Observations	536	536	536	536
Subjects	134	134	134	134
Panel B: Team-competition				
Human-framed	-17.72		-19.36**	
	[9.238]		[6.940]	
Human-likeness		-63.53		-69.41**
		[32.481]		[24.068]
Gender (female)	-24.43**	-21.64*	-7.55	-4.51
	[9.055]	[9.388]	[7.132]	[7.034]
Age	-1.30***	-1.33***	-0.85***	-0.88***
	[0.364]	[0.355]	[0.235]	[0.223]
Constant	481.24***	488.11***	84.98***	92.48***
	[15.258]	[16.513]	[11.027]	[11.484]
Block-fixed effects	✓	✓	✓	✓
Observations	528	528	528	528
Subjects	132	132	132	132

Notes: Specifications (1) and (3) report random-effects regressions on the Human-framed treatment dummy. Specifications (2) and (4) report two-stage least squares random-effects regressions in which (perceived) Human-likeness is instrumented by the Human-framed treatment. Robust standard errors clustered at the subject level are reported in square brackets. All specifications include block fixed effects. * $p < .05$ ** $p < .01$ *** $p < .001$.

Table 2: Random-effects and instrumental-variable regressions on individual and intervention scores

3.2. Perceived Human-Likeness and Instrumental Variable Estimates

The reduced-form results establish that the AI framing significantly raises intervention under both incentive schemes. A natural follow-up question is whether this effect operates through participants' perceptions of teammate identity or whether additional features of the labeling manipulation that are independent of this belief measure also contribute. We address this question using a post-experimental measure of participants' beliefs about teammate identity, which allows us to recover the component of the treatment effect that operates through this belief channel. At the end of the experiment, participants rated the

perceived nature of their teammates on a 7-point Likert-scale. We constructed a perceived *human-likeness* measure by inverse-coding this scale and normalizing it to $[0, 1]$, where higher values indicate stronger perceptions of teammates being human-like. Under fixed payment, in the Human-framed treatment, the mean perceived human-likeness is 0.366 ($SD = 0.267$), with 84.72% of participants assigning positive human-likeness. In the AI-labeled treatment, the mean perceived human-likeness is 0.169 ($SD = 0.223$), with approximately half of participants (49.21%) assigning no positive human-likeness to their teammates. A similar pattern holds under team competition: in the Human-framed treatment, mean perceived human-likeness is 0.387 ($SD = 0.225$) with 90.62% of the participants assigning positive human-likeness. In the AI-labeled treatment, the mean is 0.112 ($SD = 0.151$), with more than half of participants (52.94%) assigning no positive human-likeness to their teammates (see Table 3). The difference in perceived human-likeness between treatments is statistically significant in both payment conditions ($p < .001$, two-sided t-tests). Figure 3 in the Appendix presents the kernel density estimates of this measure for both payment conditions.

Table 3: Perceived Human-Likeness p by Treatment and Condition

	Human-framed treatment					AI-labeled treatment				
Condition	Mean (SD)	Median	= 0 (%)	> 0 (%)	Obs.	Mean (SD)	Median	= 0 (%)	> 0 (%)	Obs.
Fixed-payment	0.37 (0.27)	0.33	15.28	84.72	288	0.17 (0.22)	0.17	49.21	50.79	252
Team-competition	0.39 (0.23)	0.42	9.38	90.62	256	0.11 (0.15)	0.00	52.94	47.06	272

Notes: The table reports summary statistics for the perceived human-likeness p, normalized to the unit interval $[0, 1]$, where larger values indicate stronger perceptions that teammates were human-like. Columns "= 0 (%)" and "> 0 (%)" report the percentage of observations assigning zero and positive human-likeness, respectively.

In both conditions, the Human-framed treatment shifted beliefs toward substantially greater uncertainty, with many participants no longer clearly classifying their teammates as either human or artificial. Importantly, the Human-framed treatment did not induce a uniform belief that teammates were human, it induced uncertainty. We note that this measure captures beliefs at a single point after the task is completed and may not fully reflect the beliefs that governed behavior in each block, particularly if perceptions of teammate identity evolved over the course of the experiment. The IV estimates below should therefore be interpreted as recovering the effect of perceived identity operating through this measured belief variable, rather than as capturing all channels through which the framing may affect behavior.

Because the Human-framed treatment induced heterogeneous uncertainty rather than a uniform belief of human identity, the reduced-form treatment comparison understates the component of the effect that operates through the measured belief channel: participants in the Human-framed treatment who suspected their teammates were bots would be expected to behave more similarly to participants in the AI-labeled treatment, attenuating the estimated treatment difference. To recover the effect of perceived identity on behavior while accounting for this heterogeneity, we use the Human-framed treatment indicator as an instrument for the perceived human-likeness measure in a two-stage least squares random-effects regression (2SLS). The instrument is relevant: the first-stage F-statistic is 23.48 under fixed payment and 66.52 under team competition, well above conventional thresholds for instrument strength, and is valid by design, since treatment assignment was randomized and the framing manipulation is designed to influence participants' perceptions of teammate identity while holding underlying teammate behavior constant.

Specifications (2) and (4) in Table 2 present the IV estimates alongside the reduced-form results, allowing direct comparison. Under fixed payment (Panel A), perceived human-likeness has a significant negative effect on the intervention score ($p = .024$), confirming that the reduced-form treatment effect reflects a genuine belief channel: participants who more strongly perceived their teammates as human-like intervened significantly less in their teammates' areas of responsibility. The individual score is not significantly affected ($p = .601$). Under team competition (Panel B), the negative effect of perceived human-likeness on intervention persists and strengthens ($p = .003$). The individual score shows a directionally negative but insignificant effect ($p = .238$), consistent with the reduced-form pattern.

The comparison between reduced-form and IV estimates is itself informative: the IV estimates are larger in magnitude than the corresponding reduced-form estimates in both conditions. This amplification is consistent with the manipulation check showing that the Human-framed treatment induced imperfect and heterogeneous belief shifts. Importantly, the direction and significance of the intervention effect remain unchanged between estimators, reinforcing the robustness of the main finding.

Together, the reduced-form and IV results present a consistent pattern of evidence. Perceived teammate identity significantly shapes intervention behavior under both incentive schemes, with participants exerting substantially more compensatory effort when collaborating with teammates they perceive as artificial.

3.3. Simple model of effort allocation in hybrid teams

To organize these findings, we propose a simple model of effort allocation in hybrid teams that formalizes the role of perceived teammate identity in shaping the incentive to intervene.

A key feature of the experimental design is that the AI-labeled treatment corresponds to the objectively certain case: participants are told they are working with bots, which is precisely what they are. The Human-framed treatment, by contrast, introduces a belief distortion: participants are uncertain whether their teammates are artificial or human, even though the agents are identical to those in the AI-labeled treatment and follow the same programmed behavior. Behavioral differences across treatments therefore do not reflect differences in actual teammate performance, but purely differences in participants' beliefs about teammate identity.

We model this belief as $p \in [0, 1]$, the participant's subjective probability that a teammate is human, which corresponds to the perceived human-likeness measure p constructed from the post-experimental Likert-scale. In the AI-labeled treatment $p = 0$ with certainty, consistent with the low mean human-likeness of 0.169 and 0.112 observed in that treatment under fixed payment and competition respectively, with approximately half of participants assigning zero human-likeness to their teammates. In the Human-framed treatment the framing induces $p^* \in (0, 1]$, which is unobserved and heterogeneous across participants: the overwhelming majority of participants assign positive human-likeness (84.72% under fixed payment and 90.62% under competition), with substantial variation ($SD = 0.267$ and 0.225 respectively), confirming that the treatment induced uncertainty rather than a uniform belief of human identity.

Each participant maximises a utility function

$$U_i \;=\; \lambda \cdot W\big(Q(e_i,\, \gamma_i,\, S_A)\big) \;+\; \alpha(p) \cdot V(\gamma_i) \;-\; c(e_i + \gamma_i), \tag{2}$$

where e_i denotes effort in the participant's own segment, γ_i denotes intervention effort in the agents' segments, and S_A denotes the agents' scores, which are exogenous to the participant. $Q(\cdot)$ is the team

score, $W(Q)$ captures the utility from performing well with $W' > 0$ and $W'' < 0$, $V(\gamma_i)$ is the non-monetary benefit of helping with $V' > 0$ and $V'' < 0$, and $c(e_i + \gamma_i)$ is the effort cost with $c' > 0$ and $c'' > 0$. The scalar $\lambda > 0$ weights score utility relative to the other components.

The identity parameter $\alpha(p)$ scales the marginal utility of intervention by the participant's belief about teammate identity:

$$\alpha(p) \;=\; (1-p)\cdot\alpha_{AI} \;+\; p\cdot\alpha_{H}, \qquad \alpha_{AI} > \alpha_{H} \geq 0. \tag{3}$$

Since $\alpha'(p) = \alpha_H - \alpha_{AI} < 0$, the non-monetary benefit of helping is largest when the participant is certain the teammates are artificial ($p = 0$) and smallest when certain they are human ($p = 1$). Three channels identified in the literature support this ordering. First, algorithm aversion (Dietvorst et al., 2015) suggests that people react more strongly to errors made by algorithms than to equivalent errors made by humans, increasing the felt urgency to intervene when an AI teammate underperforms. Second, responsibility attribution (Gazit et al., 2023) implies that individuals are less willing to offload the duty of correcting poor outcomes onto an artificial agent than onto a human colleague, creating a stronger perceived obligation to monitor and intervene in an AI teammate's area. Third, social norms governing workplace interaction suppress intervention when the teammate is perceived as human, stepping into a colleague's area of responsibility risks being seen as intrusive or disrespectful, whereas this concern does not arise when the teammate is perceived as an artificial agent lacking the kind of mind that would register such social violations (Wiese et al., 2022). Each of these channels independently implies $\alpha_{AI} > \alpha_H$, and together they reinforce the assumption.

An interior optimum γ_i^* satisfies the first-order condition

$$\alpha(p)\cdot V'(\gamma_i^*) \;+\; \lambda\cdot W'(Q^*)\cdot Q_\gamma \;=\; c'(e_i+\gamma_i^*), \tag{4}$$

where $Q_\gamma \equiv \partial Q/\partial\gamma_i$. As whether intervention increases or decreases the team score depends on task-specific parameters, agent reliability, the asymmetric scoring rule, and the timing of selections, the sign of $\partial Q/\partial\gamma_i$ is not specified.

The left-hand side of (4) is the marginal benefit of intervention, comprising the belief-weighted non-monetary gain and the score-utility effect, and the right-hand side is the marginal effort cost. Define $F(\gamma_i, p)$ as the left-hand side of (4) minus the right-hand side. The second-order condition $F_\gamma < 0$ holds under the maintained concavity and convexity assumptions whenever $Q_{\gamma\gamma} \leq 0$, a regularity condition that requires only that the score is weakly concave in intervention.

Differentiating the first-order condition with respect to p and applying the implicit function theorem gives

$$\frac{\partial\gamma_i^*}{\partial p} \;=\; -\frac{F_p}{F_\gamma} \;=\; -\frac{\alpha'(p)\cdot V'(\gamma_i^*)}{F_\gamma}. \tag{5}$$

Since $\alpha'(p) = \alpha_H - \alpha_{AI} < 0$, $V'(\gamma_i^*) > 0$, and $F_\gamma < 0$, the numerator of (5) is positive and the denominator is negative, so $\partial\gamma_i^*/\partial p < 0$. Optimal intervention is strictly decreasing in the belief that the teammate is human. This result holds regardless of the sign of Q_γ, regardless of whether an own-segment effort constraint binds, and regardless of how S_A relates to γ_i.

Because the AI-labeled treatment implies $p = 0$, whereas the Human-framed treatment induces $p^* > 0$ for at least some participants, the model predicts $\gamma_i^*(0) > \gamma_i^*(p^*)$: participants should intervene more in

the AI-labeled treatment than in the Human-framed treatment, even though the actual agents and their performance are identical across treatments.

Our results are consistent with this prediction. The intervention score is significantly lower when teammates are presented with human avatars and names than when they are transparently labeled as bots, both under fixed payment and under team competition. The robustness of this finding across incentive conditions is consistent with the model's prediction that the belief-induced difference in intervention operates through the non-monetary benefit of helping, captured by $\alpha(p)$, independently of the monetary incentive environment. The model takes no stand on whether intervention raises or lowers team performance. Equally, the model does not predict how competitive incentives interact with the identity effect: the weight λ shifts the level of intervention through the first-order condition, but its interaction with the belief-induced difference $\partial\gamma_i^*/\partial p$ depends on higher-order derivatives that are not signed in general.

The mechanism requires only three ingredients: (i) participants derive non-monetary utility $V(\gamma_i)$ from intervening; (ii) this utility is weighted by $\alpha(p)$, which is strictly larger when participants believe their teammates are AI; and (iii) V is strictly increasing in γ_i, so the marginal benefit of intervention is always positive and belief-sensitive. No assumption about the relationship between intervention and team score is needed to derive this prediction.

An alternative model in which human teammates generate higher non-monetary utility from helping than artificial teammates, for example, through empathy, reciprocity, or felt social obligation, would imply $\alpha_H > \alpha_{AI}$ and therefore $\alpha'(p) > 0$. By equation (5), this would predict $\partial\gamma_i^*/\partial p > 0$: optimal intervention would be increasing in the belief that the teammate is human, yielding higher intervention scores in the Human-framed treatment than in the AI-labeled treatment. Our results uniformly contradict this prediction: intervention is significantly lower in the Human-framed treatment under both incentive schemes, and IV estimates indicate that this effect operates through the perceived human-likeness channel (Table 2), providing direct empirical support for the assumption that the non-monetary benefit of helping is higher when teammates are perceived as artificial than when they are perceived as human ($\alpha_H < \alpha_{AI}$). The boundary case $\alpha_H = \alpha_{AI}$, which would predict no treatment difference in intervention, is likewise rejected by the statistically significant effects reported in Table 2.

4. Concluding Remarks

While stylized, the experimental design captures central coordination problems that arise in many contemporary production environments. In such settings, workers are typically assigned individual areas of responsibility but may reallocate attention and effort toward teammates when underperformance or missed tasks become visible. The experiment is not intended to replicate any specific workplace, but rather to isolate the behavioral mechanisms governing effort allocation and intervention in hybrid teams.

Taken together, our findings show that cooperative behavior in hybrid teams is not solely determined by objective task performance or reliability. Instead, it is strongly shaped by beliefs about the nature of one's collaborators. By demonstrating that perceived teammate identity alters cooperative effort across incentive structures, our study highlights how AI integration can reshape team production through belief channels that operate independently of actual teammate capabilities.

Our results provide empirical support for the key assumption of the proposed model: that the non-monetary benefit of helping is higher when teammates are perceived as artificial than when they are

perceived as human. As shown in Section 3.3, an alternative model in which human teammates generate higher non-monetary utility from helping would predict higher intervention in the Human-framed treatment, which our results uniformly contradict. The robustness of the intervention effect across both incentive schemes is further consistent with the model's prediction that the belief-induced difference in cooperative effort operates through the non-monetary benefit channel, independently of the monetary incentive environment.

Two limitations of the current study point to directions for future work. First, our design elicits participants' beliefs about teammate identity through a post-experimental questionnaire, capturing beliefs at a single point in time after the task is completed. It may therefore not fully reflect the beliefs that governed behavior, particularly if perceptions of teammate identity evolved over the course of the experiment. Future studies that elicit beliefs at multiple points during the task would allow a sharper test of the belief channel and shed light on whether individual differences in sensitivity to the framing manipulation or changes in belief over time drive heterogeneity in intervention behavior.

Second, our results are consistent with several distinct mechanisms through which perceived AI identity raises the willingness to intervene, for example, a stronger reaction to algorithmic errors, a heightened sense of personal responsibility when working with AI, and the absence of social norms that would make intervention in a human colleague's area feel intrusive. The current design cannot separate these channels, as all of them are activated simultaneously by the AI-framing. Future work could disentangle them, for instance, by varying whether agent errors are made visible to participants, by eliciting responsibility attributions directly, or by manipulating whether intervention is observable to teammates.

Our findings carry a practical implication that extends beyond the laboratory. Organizations designing hybrid teams should recognize that how artificial teammates are labeled and presented shapes how human team members allocate effort and, ultimately, how productive the team is. In settings where mutual monitoring and compensatory intervention are valuable, transparently identifying collaborators as artificial agents can increase the cooperative effort. Understanding these behavioral responses is therefore critical for designing effective human–AI collaborations in modern organizations.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process.

During the preparation of this work the authors used ChatGPT and Claude in order to check for spelling and punctuation mistakes, and improve the readability of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Bornstein, G. (1992). The free-rider problem in intergroup conflicts over step-level and continuous public goods. *Journal of Personality and Social Psychology*, 62(4):597–606.

Bornstein, G. and Ben-Yossef, M. (1994). Cooperation in intergroup and single-group social dilemmas. *Journal of Experimental Social Psychology*, 30(1):52–67.

Bornstein, G. and Gneezy, U. (2002). Price competition between teams. *Experimental Economics*, 5(1):29–38.

Bornstein, G., Gneezy, U., and Nagel, R. (2002). The effect of intergroup competition on group coordination: An experimental study. *Games and Economic Behavior*, 41(1):1–25.

Che, Y.-K. and Yoo, S.-W. (2001). Optimal incentives in teams. *American Economic Review*, 91(3):525–541.

Cheng, X., Dogan, M., and Yildirim, P. (2025). Artificial intelligence in team dynamics: Who gets replaced and why? *arXiv preprint arXiv:2506.12337*.

Cioffi, R., Travaglioni, M., Piscitelli, G., Petrillo, A., and De Felice, F. (2020). Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions. *Sustainability*, 12(2):492.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–474.

Dell'Acqua, M. et al. (2025). Super mario meets ai: Experimental effects of automation and skills on team performance and coordination. *The Review of Economics and Statistics*, 107(4):951–971.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.

Englmaier, F., Grimm, S., Grothe, D., Schindler, D., and Schudy, S. (2024). The effect of incentives in nonroutine analytical team tasks. *Journal of Political Economy*, 132(8):2695–2730.

Gazit, L., Arazy, O., and Hertz, U. (2023). Choosing between human and algorithmic advisors: The role of responsibility sharing. *Computers in Human Behavior: Artificial Humans*, 1(2):100009.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, 118(3):1049–1074.

Hagemann, V., Rieth, M., Suresh, A., and Kirchner, F. (2023). Human-ai teams-challenges for a team-centered ai at work. *Frontiers in Artificial Intelligence*, 6:1252897.

Hamilton, B. H., Nickerson, J. A., and Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of political Economy*, 111(3):465–497.

Kandel, E. and Lazear, E. P. (1992). Peer pressure and partnerships. *Journal of Political Economy*, 100(4):801–817.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *Quarterly Journal of Economics*, 122(3):1067–1101.

Niszczota, P., Grzegorczyk, T., and Pastukhov, A. (2025). People are highly cooperative with large language models, especially when communication is possible or following human interaction. *arXiv preprint arXiv:2507.18639*.

Qiao, Z., Tabarrok, A., Zubrickas, R., and Cason, T. N. (2026). Norms in conflict: Why ai advisors fail to improve human coordination. Working paper.

Walliser, J. C., de Visser, E. J., Wiese, E., and Shaw, T. H. (2019). Team structure and team building improve human–machine teaming with autonomous agents. *Journal of Cognitive Engineering and Decision Making*, 13(4):258–278.

Wiese, E., Weis, P., Bigman, Y., Kapsaskis, K., and Gray, K. (2022). It's a match: Task assignment in human-collaboration depends on mind perception. *International Journal of Social Robotics*, pages 1–8.

Zercher, B., Jussupow, E., Benke, I., and Heinzl, A. (2025). How can teams benefit from AI team members? exploring the effect of generative AI on teams. *Journal of Organizational Behavior*, 46(1):1–18.

5. Appendix

5.1. Figures & Tables

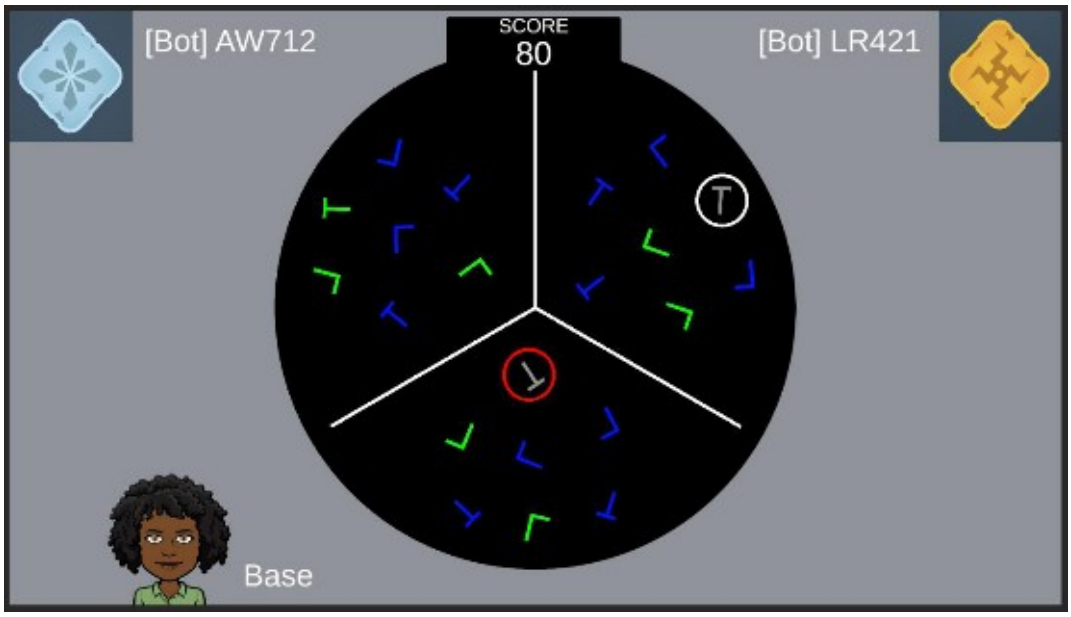

Figure 2: Screenshot of experimental user interface during the search task.

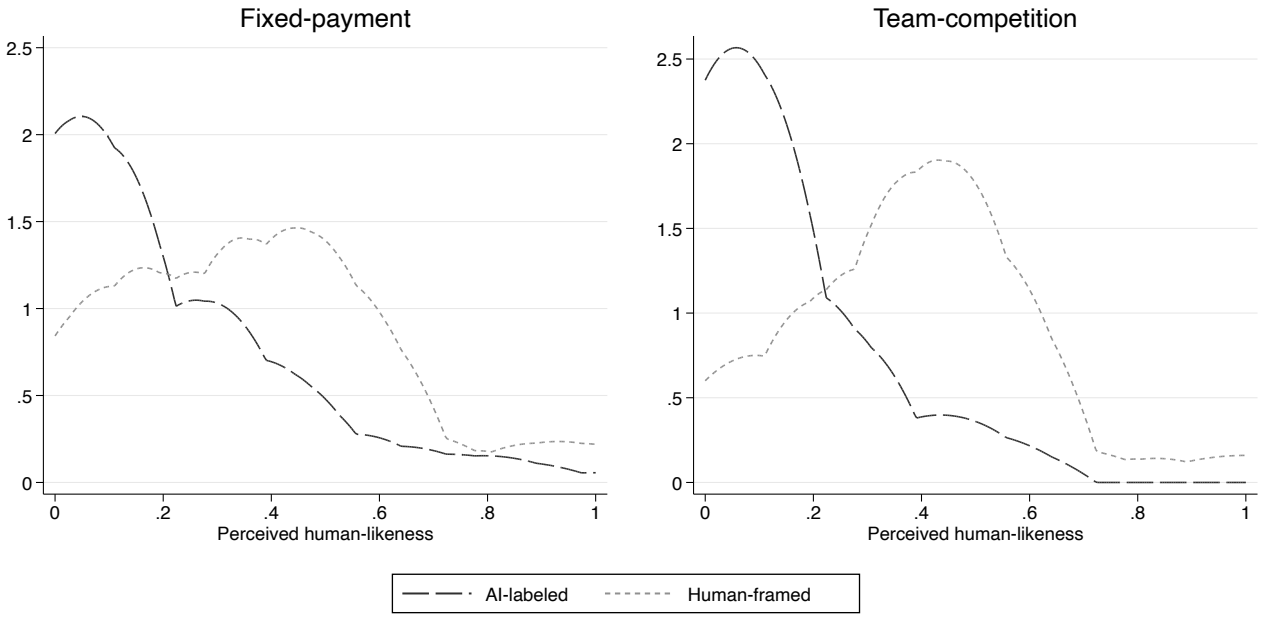

Figure 3: Kernel density estimation of perceived human-likeness between treatments and conditions.

	Individual Score		Intervention Score	
	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS
Panel A: Fixed-payment				
Human-framed	-6.02		-20.56**	
	[12.647]		[6.677]	
Human-likeness		-29.67		-101.31**
		[62.528]		[35.383]
Gender (female)	-13.80	-10.59	-9.81	1.12
	[13.239]	[13.300]	[6.608]	[7.583]
Age	-1.12	-1.10	-0.79**	-0.72*
	[0.573]	[0.588]	[0.249]	[0.304]
Constant	455.87***	458.75***	76.79***	86.60***
	[25.649]	[28.241]	[13.376]	[16.360]
First block only	✓	✓	✓	✓
Observations	134	134	134	134
Subjects	134	134	134	134
Panel B: Team-competition				
Human-framed	-28.42		-24.96**	
	[16.507]		[8.626]	
Human-likeness		-101.90		-89.52**
		[57.729]		[29.743]
Gender (female)	-12.86	-8.39	-12.24	-8.31
	[15.714]	[17.349]	[8.713]	[8.765]
Age	-2.26**	-2.30**	-1.09**	-1.13***
	[0.785]	[0.762]	[0.324]	[0.302]
Constant	518.70***	529.71***	99.49***	109.17***
	[28.095]	[31.113]	[13.800]	[14.324]
First block only	✓	✓	✓	✓
Observations	132	132	132	132
Subjects	132	132	132	132

Notes: Specifications (1) and (3) report OLS regressions on the Human-framed treatment dummy. Specifications (2) and (4) report two-stage least squares regressions in which (perceived) Human-likeness is instrumented by the Human-framed treatment. Robust standard errors are reported in square brackets. * $p < .05$ ** $p < .01$ *** $p < .001$.

Table 4: Random-effects and instrumental-variable regressions on individual and intervention scores

5.2. Instructions

Note: Text in square brackets indicates a new screen. Text in round brackets indicates form fields or pictures. Text in italics indicates differences in the Team-competition condition.

[Welcome!]
Thank you for choosing to participate in our study.
In the next few screens you will go over the consent form, answer a few questions about yourself, and then you will get started with the game.
After reading the consent form if you choose not to participate, you can do so. In that case you will receive a code to enter into Prolific to release you from the study.
We hope you enjoy the game!

[Consent Form]
What information will we store?
We store the answers and reactions to all tasks and questions, including the time it takes you to complete the tasks and the mouse movement data. We also record the date and time you conducted the study.
Data protection
We will store your anonymized Prolific ID as your subject code. A link between this subject code and your raw data is available to the test leaders, its use is strictly bound to your participation. The raw data will be stored in anonymized form and used for the qualitative and statistical analysis of their content. Accordingly, no conclusions can be drawn about the participants from the raw data. For this purpose, please contact the researcher through the Prolific site.

I hereby declare my consent to participate in the study described above.

[All about YOU!]
The following questions are related to you. There are no right or wrong answers. Please answer to the best of your ability.
What is your age? (open text field)
What is your biological sex? (Male / Female / Intersex)
What is your gender? (open text field)
Are you fluent in English? (Yes / No)

[About the game. . .]
This is a team search game you will play with two other players. You and your teammates will be working together to try and earn the most points.
If your team ranks in the top 20% of all teams, you will receive a £1 bonus.
You will be asked to pick an avatar and enter your game name.
If others are online, they will join and also choose avatars and names. If others are not online, AI bots will join indicated by an account name beginning with [Bot] and a number.
You will receive instructions and a practice round before the game begins.

[Please read the instructions carefully.]
You and 2 teammates will play a search game together. Your team will need to find target symbols, mixed with similar-looking symbols meant to distract you.
The more targets you and your teammates find together, the higher your team score will be.
On a given trial, you will see 4 different types of symbols that differ in shape and color. They can be green or blue in color and "L" or "T" in shape.

[Target Selection]
However, only ONE symbol will be your target so you should respond only to those symbols.
To respond to a target, move your mouse to the target and click your mouse.
If you select the right symbol, a white circle will appear around the target and the target will turn gray.

[Missed Targets]
During each trial there will be multiple targets for you and your teammates to find.
Unidentified targets will begin to disappear over time, and you and your teammates will lose the opportunity to respond to them.
You should respond as fast as possible so you can detect the targets before they disappear.
If you or your teammates select an incorrect symbol or miss a target, the symbol will be circled in red and turn gray.

[Board Setup]
The game board is split into 3 sections: The bottom section is assigned to you. The left side to Teammate 1. The right side to Teammate 2.

You are responsible for your section, but if you think your teammates need help, you can respond to targets in their sections too.

[Scoring]
On the top of the screen, you will see the current team score displayed.
Depending on who responds and in what section, your team can earn a different number of points.
Remember, this is a team game. / *As a reminder, this is a team game, and if your team ranks in the top 20% of all teams, you will receive a £1 bonus.*
Missing a target in any section means a loss of points for the team. Selecting targets in one's own section gives the team more points than in a teammate's section. Selecting a distractor in one's own section deducts less points for the team than in a teammate's section.

	Teammates' Sections	Your Section
Miss the Target	0 points	0 points
Correct Selection (Target)	+5 points	+10 points
Wrong Selection (Distractor)	-10 points	-5 points

[Time to Play!]
Once all players are ready, you will begin the practice round. Your team's target for this game is shown here.
Your team's target for this game is shown below. This is the target you will search for in the game to earn points.
(L or T shown)

[Time to Play!]

That was the training. Now comes the real game!

Once all players are ready, you will begin. We hope you have fun.

As a reminder, your team's target for this game is below. This is the target you will search for in the game to earn points.

(L or T shown)