Super Mario meets AI: Experimental Effects of Automation and Skills on Team Performance and Coordination *

Fabrizio Dell'Acqua* Bruce Kogut** Patryk Perkowski**

**Harvard Business School*

***Columbia Business School*

March 2, 2023

*Fabrizio Dell'Acqua (fdellacqua@hbs.edu), Bruce Kogut (bruce.kogut@columbia.edu), Patryk Perkowski (PPerkowski22@gsb.columbia.edu). This research was kindly funded by The Sanford C. Bernstein Co. Center for Leadership and Ethics at Columbia Business School and a fellowship from the Columbia Business School Behavioral Research Lab. We thank Shreya Jha, Laurel Kim, and Valeria Viazmytynova for helpful research assistance, and Ashley Culver and Yaritza Perez for setting up the experiment. For helpful feedback, we thank Ian Cockburn, Jeff Furman, Don Green, Shane Greenstein, Jorge Guzman, Ebehi Iyoha, Megan MacGarvie, Stephan Meier, Angela Ryu, Robert Seamans, and Dan Wang, and for their comments on the paper, Maria Abascal, Jordi Colomer, Daniel Keum, Amit Khandelwal and Richard Nelson. Finally, we thank seminar participants at Columbia Business School and the NBER Productivity, Innovation, and Entrepreneurship

Abstract

This article studies the effects of the introduction of artificial intelligence into teams in a laboratory experiment. We randomly assign automated co-workers into "laboratory firms" who are playing a team-based game on the Nintendo Switch console. We demonstrate that even in a task where AI outperforms humans, automation decreases overall team performance and increases coordination failures. These effects are especially large in the short-term and in low- and medium-skilled teams. Moreover, automation reduces team trust and individual effort provision. Our results support the implication that improving collaborative human-machine teams is key to the positive transformation that AI may bring to organizations.

Keywords: Artificial intelligence, teams, videogame experiments, human-machine interaction, productivity, routines, skills, tacit knowledge.
JEL Classification: J24, C92, O33, J01

Program. We thank the editor and three anonymous referees, whose feedback greatly improved the paper. The experiment in this paper was approved by Columbia University IRB (IRB-AAAS5372) and pre-registered with the AEA RCT Registry (AEARCTR-0004382). All errors are our own.

1 Introduction

Artificial intelligence (AI) is a toolkit of software algorithms that is used for learning from data. These algorithms infer patterns in the data that permit predictions of complicated phenomena, classify people by the proximity of their features in a multi-dimensional space, and improve on the statistical precision when models are over-fitted to the data.[1] AI algorithms have many applications, from facial recognition, assignment of articles to topics, identifying defects in manufacturing, enabling bots to respond to and query customers at call centers or through virtual assistants, e.g., Amazon's Alexa.

The introduction of AI into the workplace will have fundamental implications for the internal organization of firms. Prior work has examined the types of jobs that are likely to be impacted by automation and the specific tasks where machine learning outperforms humans (Brynjolfsson et al., 2018; Felten et al., 2018; Webb, 2020). In this paper, we take a complementary approach by studying the intra-organizational impacts of automation in collaborative production tasks that feature human-machine interaction. Our goal is to measure both the direct team performance consequences of automation, and also how it changes human coworker behaviors and attitudes.

A rich literature in evolutionary economics (Nelson and Winter, 1982) and the organizational sciences (e.g., March and Simon, 1958) has studied the use of routines in organizations to coordinate the actions of their members. While automation may improve performance in tasks suited to machine learning, we posit that their introduction presents a larger trade-off by disrupting these routines. Additionally, we explore the distributional consequences of introducing AI into teams. A central concern in the recent economics literature is the question of the impact of automation over the skill distribution. Many

[1] For an accessible introduction to the statistical basis of AI and machine learning, see Abu-Mostafa et al. (2012); James et al. (2013).

have speculated that AI-based automation has skill-biased consequences in line with previous generations of information technology. However, there is little direct evidence for this bias, as well as the mechanisms that are behind it.

Empirically studying these questions is difficult for a few reasons. Estimating the causal effect of automation requires clean variation in the set of firms that adopt AI. However, technology adoption decisions are endogenous to firm characteristics, and the use of technology varies widely across industry, geography, and even within the same firm. For that reason, non-experimental approaches may suffer from omitted variable bias and simultaneity, making it difficult to disentangle, for example, whether AI improves performance or whether higher-performing organizations are more likely to adopt AI. Even if we could obtain clean variation in the organizations that adopt AI, we need to collect fine-grained data before and after the intervention on team productivity and potential mechanisms, such as team coordination.

We circumvent these challenges by running a laboratory study with "experimental firms." Our approach stems from a rich tradition in experimental organizational economics that argues that we can advance our study of organizational economics by creating highly reduced-form models of more complex organizations (see Camerer and Weber, 2012 for an overview). Our experiment involves participants completing a group-based task from the "Super Mario Party" game on the Nintendo Switch console.[2] We grouped participants into firms of four, consisting of two teams of two players each. Teams must work together on a task that involves retrieving ingredients to complete recipes. Each recipe contains three ingredients, and teams have one minute per round to collect as many ingredients as possible. The task requires team coordination, communication, and strategic planning, and represents an abstract form of many real-world tasks in

[2]Per Järvelä et al. (2014)'s recommendation regarding game quality, we note that Super Mario Party has received good ratings from 7.3/10, IGN; 76%, Metacritic; and 5/5, Common Sense Media.

organizations.

The research lab design allows us to experimentally replace human players with automated coworkers to study the effect of automation on team performance and behavior. The "Super Mario Party" game has a built-in AI that outperforms the vast majority of human players. Halfway through game-play, we select a random subset of firms to receive the automated co-worker, and then a random worker within the laboratory firm to be replaced by the automated co-worker. This substitution generates exogenous variation in (i) the firms which are automated, (ii) the team that receives the AI substitution, and (iii) the specific human player who is substituted. We additionally included a control condition where a third of the teams remain the same, and a traditional turnover/new hire condition for another third of the teams, where a randomly-selected player was substituted by a human player from another firm. These experimental conditions allow us to test not only the causal effect of automation but also the extent to which automation differs from traditional employee turnover.

We classify our results as related to three literatures: the effect of automation on productivity and coordination within teams, peer effects in the workplace, and automation and skill bias. The primary experimental results concern the impact of automation on team productivity. The results indicate that the introduction of a high-performing automated player reduces team performance, as these firms return fewer ingredients. These effects are especially strong in the round following automation and remain present until the final round of play. The experimental design allows us to decompose this aggregate performance change into its team-component parts. We find that automation decreases performance not only in teams that receive the automated agent, but also in teams (housed in the same laboratory firm) who do not receive an automated agent.

Moreover, we test for how well teams coordinate among their members.[3] The productivity of teams in our task is highly dependent upon the ability of members to coordinate their actions efficiently. An innovative contribution of the study is the measure of coordination that directly affects team performance. Research assistants watched videos of gameplay and coded the number of times players bumped into one another as our measure of coordination failures. Our results show that automation leads to a sharp increase in coordination failures between humans and AI, particularly in the short-term.

Our second set of results concern peer effects, and whether automation led to changes in the beliefs and behaviors of human co-workers. Survey responses following the experiment revealed that, despite no change in financial incentives, automation led to a decrease in team trust and lower effort provision for players assigned to play with the automated agent. These results indicate that introduction of intelligent machines in firms is likely to have relevant and potentially detrimental consequences on culture and employee motivation through their effects on human workers.

Finally, we explore the distributional consequences of automation in relation to the skill makeup of teams. Prior to the start of the experiment, we collected rich skills data on all participants in the experiment. Our results indicate that the automation leads to large performance decreases in low- and medium-skilled teams, while high-skilled teams return slightly more ingredients following automation. These results indicate that the skills of teams are an important driver of the overall effects of automation, and suggest that the routines used by high-skilled teams allowed them to more easily absorb the introduction of an automated agent.

Our paper contributes to the growing literature on the effects of artificial intelligence

[3] Cohen and Bacdayan (1994) show through an innovative card game that pairs of players develop heuristic routines to guide their coordination.

on jobs and workplace tasks. Prior work in this line of research has generated predictions about the types of jobs that are likely to be impacted by artificial intelligence from the task content of jobs (Frey and Osborne, 2017; Brynjolfsson et al., 2018; Felten et al., 2018; Webb, 2020). A central focus in this research stream has been the technical suitability of tasks for machine learning (i.e., understanding the types of tasks where AI can outperform humans given technological progress and the state of machine learning). Our paper contributes to this literature by directing attention toward the organizational aspects of work and highlighting how automation may adversely impact team coordination. In our setting, automated agents outperform human ones when they play in isolation, but replacing a human player with an automated one decreases team performance and increases the difficulty of team coordination.

Second, our paper contributes to the literature on peer effects in the workplace but in the context of the effects of AI on worker productivity and behavior. Overall, research has found that worker productivity is influenced by the productivity of peers (Mas and Moretti, 2009; Falk and Ichino, 2006; Jackson and Bruegmann, 2009; De Grip and Sauermann, 2012), especially by higher-productivity peers (Mas and Moretti, 2009). Our paper examines peer effects with automated co-workers rather than human ones. Our results show that participants randomly assigned to work with a (high-productivity) AI report less effort, suggesting that the introduction of automated coworkers may have important motivational consequences on the human workers that remain.

Lastly, our paper contributes to the literature on automation and skill-bias by providing micro-level experimental evidence that the impacts of automation depend on the skills of workers. Previous work has theorized that automation is skill-biased, like previous technologies such as the Internet and computing (see Goldin and Katz, 1998), and increases labor demand toward more highly skilled workers relative to the less skilled

ones (Acemoglu and Restrepo, 2020b).[4] In our setting, we find that low- and middle-skilled teams perform worse following the introduction of an automated agent, while high-skilled teams are more successful in integrating their automated co-workers. Such causal evidence is rare in the literature, and is very relevant given the broad implications that AI will have for firms and labor markets moving forward.

The remainder of the paper proceeds as follows. Section 2 describes our pre-registered experiment, while Section 3 presents our empirical strategy. Section 4 covers the results, Section 5 provides a broader discussion of the findings, and Section 6 concludes on the societal implications of the central findings.

2 Experimental Design

The foundational design of the experiment used in our paper draws upon the behavioral economics studies by Weber and Camerer (2003) and Camerer and Weber (2008) on growing culture in a firm, and its further elaboration by Rick et al. (2007) on tacit knowledge. Common to these studies is an initial phase of play to allow teams to learn to coordinate and acquire a language or more broadly tacit knowledge held commonly among its members, followed by a second phase of play that disrupts these cultures and tacit routines by merging separate teams to cooperate on the same task. In particular, in Rick et al. (2007)'s study, teams of four complete a picture-naming task that requires participants to order pictures in the proper order. One member of the group (the "manager") has the correct order of pictures and must convey information to other team members so they identify pictures in the right order in the shortest amount of time. In the tacit-intensive condition, the player with the correct order of pictures

[4]For empirical evidence at the macroeconomic level, see Graetz and Michaels (2018) and Acemoglu and Restrepo (2020a). Meanwhile, the engineering and organizational literature has been more oriented towards complementarities achieved through human and robotics interactions. For recent examples, see the studies by Hsieh et al. (2020) and Ju (2015).

remains the same throughout all rounds, allowing the team to develop the know-how to complete the task. In the tacit-restricted condition, players alternate as the "manager" each round, which interferes with the team's ability to coordinate using tacit knowledge.

In the experiment that we designed, we adopt the same structure of one phase for learning (in our experiment, Phase 2), followed by a phase that introduces the treatments of interest (in our experiment, Phase 3). In our case, Phase 2 consists of two two-people teams that constitute a firm located in a room with a large screen to display a Super-Mario game. Each of the two teams is co-located and plays simultaneously, with incentives pegged to the organizational performance, i.e., the productivity of both teams. We also hold out a third of the teams throughout both phases of play without further treatment; these teams are high in tacit knowledge and provide a benchmark against which the AI teams and teams with new (human) hires can be compared. Moreover, as we have observations on individual skills and team performance, we are able to analyze both skills and team contributions.

Before we further discuss the details of our experimental design, two points are warranted. First, lab experiments are popular tools to study a wide range of organizational issues. These experiments require researchers to create skeletal models of more complex organizations. While lab experiments cannot capture all of the richness of modern-day firms, they provide a natural training ground to test and refine theories (Camerer and Weber, 2012), and "can help get a handle on the basic processes underlying [real-world] phenomena" (Weber and Camerer, 2003). Moreover, lab experiments are better suited to study mechanisms that can help create more generalizable knowledge occurring within organizational processes. By allowing researchers to manipulate key variables of interest systematically, lab experiments provide a natural training ground to test theories regarding organizations.

A second aspect of the experiment that deserves comment is the use of video games. While research using video games faces the criticism of running "the risk of appearing frivolous and less serious than other types of laboratory research," video games tend to be more cognitively demanding and motivating; they require more coordination, communication, and strategic planning than traditional experimental tasks (Washburn, 2003, p. 187ff.). Moreover, video games provide a natural testing group for experimenting with artificial intelligence. As computer scientists Georgios Yannakakis and Julian Togelius argue in their 2018 book, "games provide the best benchmarks for AI because of the way they are designed to challenge many different human cognitive abilities, as well as for, their technical convenience and the availability of human data" (Yannakakis and Togelius (2018), p. xiii). In particular, the video game used in this experiment resembles coordination tasks that occur in organizations and allows us to compare the performance of human versus automated agents.[5]

2.1 Task

We grouped participants into experimental firms of four players each and tasked them with playing a team-based mini-game using the "Super Mario Party" game on the Nintendo Switch console. "Super Mario Party" is a multiplayer game with 80 free-for-all and team-based mini-games that has sold over 18 million units as of December 2022.[6] We randomly assigned each participant to one of four "Super Mario Party" game characters (Mario, Peach, Luigi, and Daisy), which do not vary in their skills or tasks.

Our experimental task focused on the "Dash and Dine" mini-game. In the task, each group of four is subdivided into two teams of two players each. Participants must work with their assigned partner to compile ingredients to complete recipes. The ingredients

[5]Another relevant study is Englmaier et al. (2018), who use a real-life escape game to study the impact of different incentive structures on joint team performance in a non-routine analytical task.

[6]https://www.nintendo.co.jp/ir/en/finance/software/index.html

for each recipe are listed on the top of each team's side, and participants must retrieve these ingredients from the three tables at the bottom of the screen. Each completed recipe is worth one point, and players have one minute to complete as many recipes as possible. Figure A.3 shows a screenshot from actual gameplay. Mario and Peach are on the left team and must retrieve two tomatoes and one lettuce to complete their recipe. Meanwhile, Luigi and Daisy are on the right team and must retrieve one lettuce and two tomatoes to complete their recipe. Players cannot return ingredients to the other team's table or hand ingredients to other players.

Participants played the game using the controllers displayed in Appendix A.1. This required them to use their left thumb to toggle a joystick to move their character around the map, and their right thumb to press a button that picked up and dropped off the ingredients. No prior experience with the gaming system is necessary, and an informal survey of participants revealed that the majority of participants had no prior experience with the game. While all were able to easily learn the controls following the first phase of the experiment, we tested for and found variation in players' initial skills prior to the start of the team-level experiment.

The "Dash and Dine" mini-game has three attractive features for the purposes of this study. First, the task is coordination-based and requires both within-team and across-team (within-firm) cooperation for success. Each player must work with not only their assigned partner (to collect ingredients to complete their team's recipes) but also members of the other team (to navigate through the map without bumping into one another). Our financial bonus (described in detail below) incentivized participants to maximize total firm output (the number of ingredients returned by both teams on average), not just team output (number of ingredients returned by each team). Overall, the task captures the coordination and social skills required in real-world teams, as it

requires interdependent members to integrate their knowledge and actions in order to achieve a common goal.[7]

Second, the game includes a high-performing automated agent that can easily replace a player. While the algorithm's details are proprietary information, this game's automated agent is designed to coordinate with its partner to maximize the total number of ingredients it returns. The automated agent does many of the same actions required by humans in this game: it retrieves information from the environment (for example, seeing that their partner is heading toward a tomato and that the final remaining item to complete the recipe is lettuce), evaluates this information to make a plan of action (for example, selecting the route with the smallest distance to the lettuce), and executes this action (for example, retrieving the lettuce). We selected the most-skilled automated agent out of the three available options on the game (i.e., the computer difficulty level of "very hard"). Important for our purposes is that the automated agent in this game is high-skilled, and in fact, outperforms humans. In Figure 1, we display a kernel density plot of performance for humans versus automated agents. The automated agent returns on average 18 percent more ingredients per round than human players (7.5 ingredients for automated agents versus 6.4 ingredients for humans). This is also true throughout the vast majority of the productivity distribution. In Appendix A.4, we show that the worst performance observed for the automated agent is better than almost 30 percent of human players, while less than three percent of humans outperform the best automated agent.

Third, the task allows us to collect rich data at the firm, team, and individual level. In terms of performance data, we observe in each round the total number of ingredients collected by (i) the firm, (ii) each team, and (iii) each player. This allows us to study the

[7]Recent work documents that social skills are necessary for working well in a team (see, for example, Deming (2017) and Weidmann and Deming. (2020).

effects of automation and firm and team productivity, and individual contributions to firm and team productivity. Moreover, we supplement our productivity measures with data on coordination failures (obtained by recording the screens visible to participants) and player attitudes (through surveys in the middle and at the end of the task).

2.2 Experimental Structure

The experiment took place in the Behavioral Research Lab at Columbia Business School. Each session occurred in three phases and lasted between 45 minutes and one hour. Figure A.4 displays the general structure of the experiment.

In the first phase of the experiment, each player completed four mini-games from the Super Mario Party game.[8] We included this phase for two reasons. First, it allowed participants to get familiar with the Nintendo Switch console and controller prior to the start of the team task. Second, and more importantly, rather than rely on self-reported measures of prior experience with the console, we collect a direct measure of each participants' general skills in the game. We selected a mix of games that span visual, perceptive, and motor skills, and created a normalized index of performance in these games to use as a control in all of our regressions given its strong prognostic ability in predicting performance in the team task.

Prior to the start of the second phase, we randomized participants into one of the two team types. In some firms, participants played with the same partner for all rounds, while in others they alternated partners (see section 2.3 for additional details). Phase 2 then consisted of six rounds of the "Dash and Dine" game. Following the sixth round, participants completed a survey on team communication, trust, and effort (see Appendix A.2 for the survey questions).

Prior to the start of the third phase, we randomized teams into one of the three

[8]The games were Candy Shakedown, Barreling Along, Precision Gardening, and Follow the Money.

organizational changes. These included an automation condition, a new hire condition, and a control condition, which we further describe in section 2.3. Participants then completed another six rounds of the "Dash and Dine" game. Following the 12th round, participants completed a second survey, which contained questions on trust, effort, and AI attitudes (see Appendix A.3 for the survey questions).

As noted above, we recorded the game screen and player interactions throughout Phases 2 and 3. This allowed us to capture team scores, individual scores, and coordination failures for each round. Panel A of Table 1 displays descriptive statistics at the participant level, while Panel B of Table 1 displays them at the team-level. Our final sample contains 220 participants, split across 55 firms (or 110 teams). Given each team completed twelve rounds of gameplay, our dataset contains 1320 observations at the team-round level (110 teams * 12 rounds).

Finally, we incentivized participants by tying their performance to actual pay-offs at the firm level. Each participant $i \in \{1, 2, 3, 4\}$ in firm f received a total payout of $P_{i,f} = \$5 + \frac{\sum_{r=1}^{12} \frac{Y_{f,r}}{2}}{12}$, where r indexes rounds, and $Y_{f,r}$ measures the total number of recipes returned by firm f in round r. This payment includes a \$5 show-up fee plus a bonus based on their firm's average performance in each team across all twelve rounds. We tied the financial bonus to performance at the organizational level to incentivize cooperation between all four members of the organizations.[9]

2.3 Experimental manipulations

Our experiment followed a 2x3 factorial design where we manipulated team type (two treatment arms) and organizational change (three treatment arms). We describe each

[9]Pre-testing revealed that incentives at the team level increased competition between teams within the same firm and led some participants to block players in the other team. Incentives at the organizational level supported cooperation between participants.

manipulation below.

Team type We first randomly assigned each group of participants into one of two team types prior to the start of Phase 2. We display these manipulations in the top half of Figure 2. While each firm contained four players and each player played six rounds, we manipulated whether the participants played with the same partner or a different partner from their firm in each round. In the tacit (same-partner) condition, each player was randomly assigned a partner from their firm in round 1 and played with this partner for six rounds. In the explicit (different-partner) condition, players in the firm alternated partners every round. For example, the green player in Figure 2 played with the yellow player in round 1, the blue player in round 2, the purple player in round 3, and then back to the yellow player in round 4. Otherwise, players in both conditions played six rounds of the game with the same overall firm of four players.

Our goal with the team type manipulation was to induce differences in the opportunity for our laboratory firms to form tacit knowledge and organizational routines. The organizational learning literature has found that the development of team routines depends on interactions between members that are both repeated and near-identical (Cohen and Bacdayan, 1994; Gersick and Hackman, 1990). Repetition enhances shared knowledge among team members, thereby improving coordination. Our manipulation of randomizing whether participants retain partners or switch alters task similarity between rounds and, consequently, the opportunity to accrue tacit knowledge required for the formation of routines (Rick et al., 2007).

We included three questions in our Phase 2 survey to measure the extent to which teams coordinate their actions using explicit knowledge concerning trust, individual effort, and attitudes toward AI.

Organizational change We also assigned each firm to one of three organizational change conditions prior to the start of Phase 3. We display this manipulation in the bottom half of Figure 2. In the "automation" condition, we selected a player at random and replaced them with the automated agent described in section 2.1.[10] In the "new hire" condition, we selected a player at random and replaced them with another human player. The "new hires" completed Phases 1 and 2 at the same time but in a separate room. This was done to ensure that the "new hires" had the same experience with the game, so the only difference would be that they are now completing the task with a different firm. This also allowed us to collect skills data for these players before they joined a new firm.[11] Finally, in the "no change" condition, we kept the firm the same as in Phase 2– those in the tacit condition continued to play with the same partner, while those in the explicit condition continued to alternate partners as in the first six rounds.

3 Empirical strategy

Our design allows us to measure numerous treatment effects from both our team type and organizational change manipulations. We describe each below.

3.1 Impact of the team type manipulation

In order to estimate the effects of the team type manipulation, we run the following regression:

$$y_{t,f,r} = \alpha_0 + \alpha_1 * SamePartner_f + \theta_r + \chi_t + \epsilon_{t,f,r} \quad (1)$$

[10]The laid-off player received the show-up fee and performance bonus for Phase 2, and then left the room prior to the start of round seven. The player who replaced the laid-off player received the same financial incentives as all other players to limit pro-social concerns discussed in Brown et al. (2015).

[11]The "new hires" on average have the same performance on the skills task in Phase 1 as do the automated agents (0.37 for the AI vs 0.33 for the new hires).

where t indexes teams of two, f indexes firms of four (two teams per firm), and r indexes rounds. $SamePartner_f$ is a binary indicator that equals 1 if firm f was assigned to the same-partner (tacit) condition and 0 if they were assigned to the different-partner (explicit) condition. θ_r is a vector of round fixed effects, χ_t is a set of team control variables[12], and $\epsilon_{t,f,r}$ is the error term. α_1 captures the effect of the tacit manipulation on performance. We estimate equation 1 using robust standard errors that are clustered at the firm level.[13]

3.2 Impact of the organizational change manipulation

Total effect Our design allows us to estimate the total effect of the organizational change on firm performance, which is the difference in performance for firms of four assigned to various organizational changes. To do so, we run the following regression at the team level:[14]

$$y_{t,f,r} = \beta_0 + \beta_1 * AI_f + \beta_2 * NewHire_f + \theta_r + \chi_t + \epsilon_{t,f,r} \tag{2}$$

where t indexes teams of two, f indexes firms of four, and r indexes rounds. AI_f is a binary indicator that equals 1 if firm f was assigned to the AI organizational change condition and 0 otherwise, while $NewHire_f$ is a binary indicator that equals 1 if firm f was assigned to the new hire organizational change condition and 0 otherwise.

[12] Although our skills data from Phase 1 is collected at the individual level, we estimate equation 1 at the team-level. This requires we aggregate the individual skills data from four mini-games to the team-level. To do so, we create a z-score of average performance on each of these four games per player, and then control for the average and minimum skill of each team in our regressions.

[13] We cluster at the firm level because that is the level of random assignment (Abadie et al., 2017).

[14] Since each firm contains two teams, the firm-level impact of the organizational change is twice the average team-level treatment effect. We could theoretically estimate these treatment effects at the firm-level, but this reduces statistical precision in our estimates.

θ_r is a vector of round fixed effects, χ_t is a set of team control variables[15], and $\epsilon_{t,f,r}$ is the error term. β_1 captures the effect of automation while β_2 captures the effect of hiring a new employer on team performance. We can estimate treatment effects by team-structure condition by running equation 2 for the tacit and explicit groups separately, or estimate the overall treatment effect across both team types by pooling both conditions and including $SamePartner_f$ in the vector of controls χ_t. We estimate equation 2 using robust standard errors clustered at the firm level.

Direct and spillover effects We can also decompose the total effect of automation into its component parts. Each firm consists of two teams, and only one team receives the automated agent. This allows us to estimate both the direct effect of automation (the effect on teams who receive an automated agent) and the spillover effect (the effect on teams who do not receive the automated agent but are in a firm where the other team received the automated agent). By construction, the average total effect of automation is the average of the direct and spillover effects.

In order to estimate these effects, we replace the variables AI_f and $NewHire_f$ in equation 2 with variables $AI_{t,f}^{Direct}$ and $AI_{t,f}^{Spillover}$, and $NewHire_{t,f}^{Direct}$ and $NewHire_{t,f}^{Spillover}$, respectively. $AI_{t,f}^{Direct}$ is a binary indicator equal to 1 if team t in firm f has an automated agent on it and firm f was assigned to the automation condition, and 0 otherwise. $AI_{t,f}^{Spillover}$ is a binary indicator that equals 1 if team t in firm f does not have an automated agent in it, but firm f was assigned to the automation condition, and 0 otherwise. We define $NewHire_{t,f}^{Direct}$ and $NewHire_{t,f}^{Spillover}$ similarly. The regression

[15] χ_t contains the same controls as in equation 1 plus average pair performance in Phase 2 since this is prior to random assignment of the organizational change condition and is highly prognostic of performance in Phase 3 (Gerber and Green (2012), chapter 4).

we run to test for direct and spillover effects is thus:

$$y_{t,f,r} = \alpha_0 + \alpha_1 * AI_{t,f}^{Direct} + \alpha_2 * AI_{t,f}^{Spillover} + \alpha_3 * NewHire_{t,f}^{Direct} + \alpha_4 * NewHire_{t,f}^{Spillover} + \theta_r + \chi_t + \epsilon_{t,f,r} \quad (3)$$

In equation 3, α_1 captures the effect of automation on the pair that includes an automated agent while α_2 captures the effect of automation on the pair that does not include an automated agent but needs to coordinate with the pair that includes the automated agent. Meanwhile, α_3 captures the effect of turnover on the pair that receives a new hire while α_4 captures the effect of turnover on the pair that does not receive a new hire but needs to coordinate with the new hire. As in equation 2, we can estimate these effects across our two team types or pool them, and we use robust standard errors clustered at the firm level.

4 Results

4.1 Impacts on team performance and coordination

We begin by examining the performance over time of firms in the control condition. As discussed earlier, after Phase 2, we held out one-third of the experimental firms that did not incur a disruption to the accrual of tacit knowledge and routines developed in the first 6 rounds. We examine their performance in Panel A of Figure 3. The figure breaks out control firms by their team type condition. Those in the tacit condition played with the same partner every round, while those in the explicit condition switched partners every round. The figure illustrates that as time goes on, firms in the tacit condition begin to outperform those in the explicit condition.[16] We attribute this to

[16]If we compare performance only in Phase 3 (the final six rounds), teams in the tacit condition outperform those in the control condition ($p = 0.07$). To obtain this p-value, we limited our sample to

the development of tacit knowledge (Weber and Camerer, 2003; Rick et al., 2007) and organizational routines (Cohen and Bacdayan, 1994). Firms in this condition provide one counterfactual that we will use to examine the effects of automation.

We next compare the aggregate impact of automation and new hires on team performance. Panel B of Figure 3 plots differences in ingredients returned between AI and new hire firms compared to control firms across all twelve rounds in the experiment. The dashed line between rounds six and seven signifies when the organizational change manipulation occurred. If randomization were successful, we would not observe differences between AI and new hire groups relative to the control prior to round seven. This is indeed what we see, suggesting that assignment into the AI and new hire conditions was independent of performance in phase two.[17] Nonetheless, in order to ensure that pre-randomization differences do not drive our AI results, we control for average Phase 2 (rounds 1–6) performance when estimating treatment effects in Phase 3 (rounds 7–12). This additionally greatly reduces noise in our estimates given Phase 2 performance is highly prognostic of Phase 3 performance (and was measured prior to randomization).

The results in Figure 3 show that introducing an automated agent reduced overall team performance although, as discussed above, these automated agents outperform humans throughout the productivity distribution. We find that six rounds after the introduction of AI, these teams returned three fewer ingredients in total compared to the control team. Additionally, these results were significant for our participants, as

control (non-AI, non-new hire) units in Phase 3 and ran a regression of total ingredients on the tacit identifier, our vector of Phase 3 controls, and round indicators, with robust standard errors clustered at the team level.

[17] One slight exception is round six, where the new hire condition slightly outperforms the control group. However, with 12 overall comparisons (AI vs control and new hire vs control, for six rounds each), we would expect that around 1.2 ($= 0.10 * 12$) comparisons would return a p-value of $p < 0.10$ by random chance.

they led to an average decrease in their performance pay compensation bonus of about 13 percent for the AI group.

As a guide to interpreting the average effects estimated in the regressions discussed next, these plots lead to an important inference about the control group's characteristics. In comparison to our baseline group that did not incur a disruption to the accrual of tacit knowledge and routines, firms in the Phase 3 condition of replacing a team member (by an AI agent or by a new hire) on average performed worse than firms that were not treated by organizational change. As a result, the average effects of organizational change (the treatment) is always negative. These plots supply a verification that the firms had learned to coordinate over the first 6 rounds that were sustained in the control group through Phase 3.

4.1.1 Impact of AI on performance: direct and spillover effects

Table 2 examines the effect of automation and new hires on the total number of ingredients returned per round. Columns 1–4 display the results of Equation 2 while columns 5–8 display the results of Equation 3. These regressions pool both team type manipulations by including a binary indicator for assignment to the tacit condition, so the coefficient returned on AI and new hire represents a weighted average of the treatment effects across both team types.[18] We display the effect across all six rounds in Phase 3 in columns 1 and 5, and then break up the effect across various rounds

[18]We do this for three reasons. First, Phase 2 results indicate that the tacit manipulation did not impact team performance or use of explicit communication (see Appendix B.1). Second, Phase 3 results indicate that the effects of automation and turnover do not vary by team type (see Section B.3). Third, Columbia University's IRB closed all non-essential in-person lab studies due to COVID-19 precautions, forcing us to halt data collection on March 12, 2020. At this point, we had finished collecting data on 55/60 teams. The remaining groups were all in the explicit - new hire condition, so these estimates would be under-powered. Pooling both team types increases the precision in our estimates of the impact of automation, though we disaggregate results by team type in Section B.3.

in the remaining columns. Columns 2 and 6 examine the impact in the short-term (round 7, the round right after the organizational manipulation), columns 3 and 7 in the medium-term (rounds 8–11), and columns 4 and 8 in the final round of the task.

The results in Table 2 indicate that the introduction of high-performing automated agents decreased firm performance. Across all six rounds in Phase 3, teams in firms assigned to the AI condition returned 0.51 fewer ingredients on average per round (column 1). Each firm in our experiment consisted of two teams, so firms assigned to the automation condition returned on average 6.12 (=2 teams * 0.51 ingredients * 6 rounds) fewer ingredients in total than those in the control group. This performance difference had a meaningful impact on financial payoffs, as the bonus tied to performance was 13 percent lower.

This average effect of automation conceals important heterogeneity across time. Table 2 indicates that the negative effect of automation was highest in the round following the introduction. Column 2 in Table 2 shows that teams in firms assigned to the AI condition returned one fewer ingredient on average in round 7. This effect is economically meaningful: it represents an almost 8 percent (= 1.01/13.3) decrease in ingredients returned relative to the control mean in round 7. This negative performance effect for the teams assigned to the AI condition shrinks by half across rounds 8–11 (column 3) but is still statistically significant at conventional levels. In the medium term, firms assigned to AI improve their productivity but still underperform relative to control firms. Meanwhile, the final round of gameplay shows no difference between teams with automated agents and those in the control group (column 4). A speculative explanation is that teams assigned to the AI condition have learned how to play with the AI agent by the final round.

Our experimental design allows us to decompose the aggregate firm performance

change into its component parts. Recall that each firm consists of two teams, and only one of the teams in the firm receives an automated agent that replaces a human. We refer to this team as the "AI direct effect" and the non-changed team as the "AI spillover effect." Columns 5–8 of Table 2 examine which team is responsible for the change in firm performance. Column 5 illustrates that across Phase 3, the change in firm performance is driven by teams that receive an automated agent, who retrieve on average 0.80 fewer ingredients per round. In the immediate term, however, firm performance decreases in both directly-affected teams and spillover teams (column 6). In fact, each of these teams contributes an equal share to the firm's performance decrease.[19] This suggests a peculiarity of the behavioral response to AI, as the initial disruption affects the spillover team, while this does not happen when a new human hire is introduced. Performance in the spillover teams rebounds quickly and is not different than control teams in rounds 8–11 (column 7). However, the team with an automated agent experiences a more prolonged period of lower performance, as the coefficient only decreases by less than half from round 7 to round 12.

Our previous discussion estimates the impact of automation relative to a control group that experiences no change in its members or structure. This comparison may conflate the effect of automation with the effects of any organizational change to the experimental firm. In order to more fully understand how automation is unique in this regard, we can compare the point estimates of automation versus those of a common organizational change: traditional (human) turnover. The bottom of Table 2 contains p-values from these tests across each column. The results indicate that both automation and introducing a new human hire led to inferior performance in the firm when compared to the control condition. This is in line with our expectations, as the

[19]The point estimate on the spillover effect is slightly larger than that on the direct effect, though this difference is not statistically significant at conventional levels (p=0.83, not displayed).

control condition has built organizational routines and tacit knowledge, which are not disrupted by the introduction of a new (either human or automated) player. Looking at the point estimates, they are larger for new hires (i.e., the impact on firm performance is more negative) than the ones for automation, but this comparison is not statistically significant in any of the total effect results in columns 1–4.

Our results indicate, however, that automation and new hires operate through different organizational channels. While automation decreases performance in directly-affected and spillover teams in the short-term, the introduction of a new hire only changes performance in the directly-affected team. The change in performance in directly-affected teams is twice as large for the new hire condition compared to the automation condition ($p = 0.012$). These results highlight that the organizational implications of automation are distinct from those of traditional (human) turnover and involve relevant behavioral adjustments for all affected players.

4.1.2 Impact of AI on coordination failures

We next explore how automation influenced the ability of firms to coordinate their actions. Successful coordination in this task requires that participants (1) collect non-duplicate recipes when only one was needed, and (2) navigate the map without bumping into one another. Cases of the former were very rare in our setting, and so we focus on the number of bumps per round. This measure was coded up by research assistant who were blind to the treatment conditions. Table 3 examines the effects of the organizational change manipulations on coordination failures.

The results in Table 3 indicate that the introduction of an automated agent hampered team coordination. Teams in firms assigned to the automation condition experienced on average 1.28 more coordination failures per round across Phase 3 compared to control teams. Moreover, the relative patterns in treatment effects across rounds in columns

2–4 line up with the performance effects observed in Table 2. The effects were largest in the round following automation; AI-firms experienced 2.51 more coordination failures in round 7, before leveling off in the final round of play.

The results in columns 5–8 decompose the total effect into direct and spillover effects and show that the effect is concentrated in teams that receive an automated agent. These teams experience on average 2.81 more coordination failures across all rounds, and 3.80 more in round 7. Coordination failures weakly increase for spillover teams in round 7; although the coefficient is not statistically significant due to a larger standard error, the magnitudes across rounds follow the patterns in Table 2.

These coordination failures represent a unique consideration for automation. The results in Table 3 indicate that teams in firms that receive a new human hire do not experience more coordination failures than control firms. Moreover, the p-values at the bottom of Table 3 indicate that the increase in coordination failures is unique to automation. These results shed light on one reason why introducing high-performing automated agents decreases firm performance: automation increases the difficulty of team coordination, especially for human-AI teams.

4.2 Impact on trust, effort, and AI attitudes

In this section, we explore the impact on participant attitudes and beliefs. Following the final round of play, participants completed a 9-item survey containing questions about trust, individual effort, and attitudes toward AI.[20] We choose these three concepts because they highlight plausible mechanisms through which automation can affect organizational outcomes.

We first theorized that automation may impact organizational culture by changing team trust. Trust is an essential component of corporate culture and is linked to

[20] We list the survey questions in Appendix A.3.

improved productivity within firms by increasing decentralization (Bloom et al., 2012) and specialization (Meier et al., 2019). A growing literature in the social sciences has documented that humans do not trust AI, even if these systems outperform humans (Dietvorst et al., 2015). If humans do not trust automated agents relative to human partners, they may spend more time and attention monitoring the actions of their automated counterparts. The degree of trust thus represents potentially one way that automation compromises effective organizational routines.

A second avenue is through effort provision. The financial incentives in our experiment did not change between Phases 2 and 3 of the experiment. However, automation may decrease effort provision due to a change in intrinsic motivation if participants have a preference for completing the task with other humans. Most human players are conditional players: their effort depends on those around them. For example, Mas and Moretti (2009) find that worker productivity increases when working alongside a highly-productive (human) co-worker. A central question in this experiment is whether this is true for automated workers. If humans have a preference for working with other human players, partnering with an automated partner may reduce their intrinsic motivation and effort on the task.[21]

Finally, we explore whether human interactions with automated agents impact attitudes towards AI. A growing literature in psychology and behavioral economics discusses algorithmic aversion (Dietvorst et al., 2015). Humans feel uncomfortable with machines, are less tolerant of mistakes made by machines (Dietvorst et al., 2015), and are less likely to incorporate assistance from machines (Luo et al., 2019). Overall, these elements may hinder collaboration between humans and AI. Our survey questions here intended to measure whether subjects that interacted with AI as a partner improved or worsened

[21] In related work, Dell'Acqua (2022) finds that humans matched with better-performing AI exert less effort than those matched with lower-performing AI.

their attitudes towards AI.

Table 4 examines how automation influenced these outcomes. It displays the results of equation 3 with an additional binary indicator ("New hire, actual") that equals one if the survey respondent is the new team member to join the group, and zero otherwise. Column 1 displays the results for trust in their partner, column 2 for individual effort, and column 3 for AI attitudes. Each outcome family (trust, effort, and AI attitudes) contained three questions, which we normalize and average to create three standardized indices with mean 0 and standard deviation 1. We display the treatment effects on individual survey items in Appendix B.6.

Our results indicate that automation led to changes in components of organizational culture and individual effort provision. The results in Table 4 indicate that automation decreased trust and effort provision for directly-affected players. Players who were assigned to play with the automated agent report 1.44 standard deviations lower on our trust index. This effect is larger for the general trust question and trust in effort provision, while not statistically different from zero for trust in skills (see Appendix B.6.) Players assigned to play with the automated agent also report 0.81 standard deviations lower on our effort index, and participants were especially more likely to report that they did not pay attention (see Appendix B.6.)[22] Meanwhile, we see no effect on participants in spillover teams or in the new hire condition.[23]

[22]One potential mechanism that may help explain our findings is prosocial concerns. Brown et al. (2015) finds that replacing workers with lower-wage workers can decrease the effort of remaining players due to pro-social concerns. We believe this is unlikely in the new hire condition because the wages offered to the new hire had the same structure as the other players, and they experienced no change in effort following the new hire manipulation. If prosocial concerns drove the automation results, we would expect to observe a negative impact for all participants in the AI firm. However, the negative effort effects only occur for the player that worked with the automated agent in the last round.

[23]In Appendix B.5, we rerun these tables by interacting treatment with an indicator for the tacit

Although automation shifts trust and effort for directly-affected players, we observe no change in attitudes towards AI. Participants assigned to play directly with the AI show no difference on our AI attitudes index (0.07 standard deviations lower, p=0.81). This suggests that our results are not driven by algorithmic aversion. Instead, we theorize that our participants derive a benefit from playing with other humans.

Although the effect of automation dissipates by the final round of play, our results indicate that it led to longer-term changes in components of organizational culture and individual effort provision. The survey results presented in this section reveal that automation led to a decrease in team trust, an essential component of corporate culture, and effort provision. These results indicate that the introduction of intelligent machines in firms likely has important consequences on organizational culture and employee motivation. We return to this observation in our discussion.

4.3 Skills

The setup of our experiment allows us to also investigate how skills interact with automation to impact team productivity. As described in section 2.2, we collected rich skills data on all participants. We create a normalized z-score skill measure for each participant, and provide evidence in Appendix B.8 that this measure is strongly correlated with performance. Two features of this measure provide additional benefits for our investigation of skills and automation. First, rather than rely on broad worker categories such as wages or education as is common in the literature, we are able to directly collect each player's skills from their Phase 1 play. Second, the exogenous assignment of players to teams and teams to automation allows for a more nuanced investigation of the interaction between skills and automation. Our experimental set-up allows us to test for heterogeneity by both the skills of the replaced player and of the remaining players to uncover what is driving skill complementarity in our setting.

condition, and find limited differences in the treatment effects across the two team types.

Panel A of Table 5 examines differences in treatment effects across the skill distribution. Columns 1–3 display the effects of automation and new hire depending on the skill of the player replaced while columns 4–6 do so for the skills of the three retained players. We partition our sample by the skill distributions of the players that are replaced, and of the players that are retained.[24] Low refers to players and teams in the bottom third of the skill distribution, medium in the middle third, and high in the topthird. Column 3, for example, examines the effect of automation and new hires when a high-skilled player is removed from a team. Our key statistic to test for skill bias in our experiment is a p-value obtained from seemingly unrelated estimation, under the null hypothesis that automation has the same treatment effect across the skill distribution. We display this p-value at the bottom of Table 5.

The results in Table 5 provide experimental support for the theory that automation is skill-biased. The complementarity between automation and skills that we observe is due to the *skill-makeup of the team of remaining players*. A joint-test of equality across columns 4–6 indicates that the returns to automation vary by the skills of the remaining players ($p = 0.019$). Low- and medium- skilled teams struggle to integrate their automated co-workers and experience large performance decreases, while high-skilled teams actually return slightly more ingredients (though this effect is not statistically significant at conventional levels). On the other hand, automation does not depend on the skills of the departing player. Teams who lose a high-skilled player experience the same performance impact as those who lose a low- or medium-skilled one ($p = 0.724$). Overall, the results in Table 5 support the argument that automation and team skills are complements[25], and that the skill-bias arises not from the skills of an individual

[24]For analyzing retained human players, we take the average skill level of the three retained players.

[25]In Appendix B.9.1, we run an alternative test for complementarities using Milgrom and Roberts (1990) and Brynjolfsson and Milgrom (2013) and find further support for these results.

player, but from the skills of the other members of the team.

In Appendix B.9, we further examine what is behind the complementarity between automation and team skills documented above. We find no evidence that the complementarity is due to the automated agent returning more ingredients when matched to a high-skilled team (Appendix B.9.2). Instead, the source of complementarity stems from an improvement in the performance of high-skilled teams when assigned an automated agent versus a human control player (Appendix B.9.3). We lack the data to identify a specific mechanism through which automation improves the performance of high-skilled teams, but we can rule out explanations regarding learning (Appendix B.10), coordination failures (Appendix B.11), and exclusively direct effects (Appendix B.11.1).

Finally, to improve insight into sources of variation due to skills in our experiment, we decompose productivity by the contributions of higher- and lower-skilled players on a given team.[26] This analysis serves to distinguish between two types of production functions: one where output is a "weak link" production function of the minimal skilled worker (as analyzed in Becker and Murphy (1992)) and the second is a function of the maximally skilled worker. Panel B of Table 5 reports this analysis to see which is more predictive of a given team's performance. The results indicate that both skills matter: teams do better on average when the skills of their best and worst players improve.

4.4 Additional results

In Appendix B.3, we examining heterogeneity in the effect of automation by team type and task difficulty. We find limited evidence that the effects of automation differed by team type. Although automation has a larger (more negative) effect on tacit teams, this difference is not statistically significant (p=0.55). Meanwhile, we find that the

[26]This decomposition permits us to address the analysis in Ahmadpoor and Jones (2019) regarding the matching of skills of members to a team and the effects on productivity.

effects of automation do not vary by the difficulty of the task ($p = 0.40$).

5 Discussion

The evidence that AI augments the productivity of workers has motivated many studies about the possibilities and limits of AI augmentation (Raj and Seamans, 2019; Brynjolfsson et al., 2021). Our study provides further insights on the behavioral and organizational changes that this entails, as well as how these impacts vary with the skills of teams. Because of our experimental design, we can test whether automation is skill-biased with micro-level measurements of skills, connecting our insights to a large macroeconomic literature on skill-biased technical change (SBTC) (Acemoglu and Autor, 2010; Goos et al., 2014; Autor, 2015; Acemoglu and Restrepo, 2020b). There are, of course, limitations to connecting our set-up to the larger SBTC debate. In our experiment, all workers do the same task, so any skill bias that arises occurs within a given task that both humans and machines do. However, automation can also shift the types of tasks that are assigned to humans versus machines. Such across-task reallocation is key to macroeconomic approaches examining SBTC.

The task in our experiment requires a mixture of motor and cognitive skills, as well as the social skills for working well in a team (Deming, 2017; Weidmann and Deming., 2020). The advantage to our laboratory experiment is the ability to directly measure skills without relying on broad census classifications linking skill to occupations or tasks. Using our skills measure, Section 4.3 provides an analysis of skill-bias in automation. Our experiment found a more successful pairing of machines and humans if the human workers are highly skilled. Low- and medium-skilled teams are unable to integrate both humans and their AI avatar partners within the total organization.

Our experimental results illustrate that there are motivational costs arising from human-AI collaboration, particularly for the participant playing directly with the AI

agent. One way to understand this decrease in motivation is that the introduction of AI generates a loss in sociality in the workplace. This speculation is grounded in the questionnaire responses we reported earlier. Field experiments using dictator games suggest that sociality is a central mechanism influencing the degree of equitable sharing of monetary awards (Henrich et al., 2001, 2004). After the end of Phase 3 in our experiment, we surveyed participants and asked them to rate their partner preference from 1 (strongly prefer AI) to 7 (strongly prefer human). We display the histogram of responses in Figure B.10. Human participants prefer working with humans than with AI; 84 percent reported they prefer or strongly prefer human coworkers. Additionally, the retained human workers exert less effort when paired in a team with AI. These results underscore the conditional nature of human effort on the preference for human sociality. This non-pecuniary preference affects productivity as the presence of algorithmic co-workers negatively influences the provision of effort.

Coordination failures in our experiment provide another measure of the organizational consequences of automation. The control condition of human-human teams relying on accrued tacit routines evidenced higher productivity and coordination throughout Phase 3 of the game (Nelson and Winter, 1982; Kogut and Zander, 1992; Cohen and Bacdayan, 1994). The results point to the inference that disrupting the organization of work within and between teams is costly. As found in prior studies such as Bresnahan et al. (2002), work practices anchored in human routines and technology are complementary, and deviations of one from the other are costly. In our experiment, we were able to test this proposition directly by randomizing the assignment of players, and thereby teams, to the AI treatment introduced in Phase 3.

These observations permit speculation into the AI "productivity paradox" that the productivity gains to the adoption of AI technologies are not evident in the macro

data (see, for example, Brynjolfsson et al. (2019)). Our experimental results suggest one possible explanation may lie in the heterogeneous composition of teams in regard to human skills and in the behavioral motivations of players. The results from the experiments analyzed above point to the relevance of taking an organizational and team perspective on the relationship of the sociality found in work teams and the motivational degradation of human productivity caused by automation.

6 Conclusion

Much of the debate around the future of work has focused primarily on the types of tasks that will be substituted by AI. Our results indicate that progress in the future design of work will be determined by the routines by which teams and organizations coordinate their actions, and the design of the algorithms that support human-machine collaboration. Additionally, human responses will be fundamental. Successful AI introduction will require augmenting humans through AI rather than potentially demoralizing them through AI collaboration they do not enjoy.

One way to understand our results is that humans prefer working with humans, and AI agents play well alone. AI agents cannot be assumed to make for great team members. A challenge for the design of high-performing human and machine teams is the creation of systems of humans and machines that benefit from algorithmic learning while maintaining a motivated human workforce.

The experimental approach we use has advantages in manipulating worker environments to evaluate the causal impact of automation on team performance and worker attitudes and behaviors. The laboratory design using pairs of teams created in a laboratory setting focused on key features in current theories and discussions of the impact of AI on the future of work, not only in terms of productivity but also of coordination and motivations. We chose a coordination-based task with a highly skilled automated agent,

and selected a subset of teams at random for substituting a human worker by this AI agent. Our results indicate that the introduction of these automated agents reduced team performance, especially in low- and medium-skilled teams. We provide evidence that automation decreases team coordination, trust, and individual effort.

These reflections on the organizational and motivational consequences of automation through AI echo an older history of research on the close connection of social and technical change. The 1951 study by Eric Trist and Ken Bamforth of the adoption of a new organizational technology called the "the longwall method of coal-getting" became a classic study of organizational change as a "socio-technical" system (Trist and Bamforth, 1951). The study reported on the effects of technical change on worker stress in the mines as well as at home due to the reversal of status and pay that also affected worker motivation. Our results indicate that arriving at an understanding of the economic performance of human and AI mixed teams requires a healthy appreciation of human motivation and sociality in the changing workplaces of the future.

References

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge, "When Should You Adjust Standard Errors for Clustering?," *NBER Working Paper #24003*, 2017.

Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin, *Learning from Data*, Vol. 4, AMLBook, 2012.

Acemoglu, Daron and David Autor, "Skills, Tasks and Technologies: Implications for Employment and Earnings," Working Paper 16082, National Bureau of Economic Research June 2010.

_ **and Pascual Restrepo**, "Robots and Jobs: Evidence from US Labor Markets," *Journal of Political Economy*, June 2020, *128* (1), 2188–244.

_ **and** _ , "Unpacking Skill Bias: Automation and New Tasks," *AEA Papers and Proceedings*, May 2020, *110*, 356–61.

Ahmadpoor, Mohammad and Benjamin F. Jones, "Decoding team and individual impact in science and invention," *Proceedings of the National Academy of Sciences*, 2019, *116* (28), 13885–13890.

Autor, David, "Why are there still so many jobs? The history and future of workplace automation," *Journal of economic perspectives*, 2015, *29* (3), 3–30.

Becker, Gary S. and Kevin M. Murphy, "The division of labor, coordination costs, and knowledge," *The Quarterly journal of economics*, 1992, *107* (4), 1137–1160.

Bloom, Nicholas, Raffaella Sadun, and John Van Reenen, "The Organization of Firms Across Countries," *Quarterly Journal of Economics*, 2012, *127* (4), 1663–1706.

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt, "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence," *The Quarterly Journal of Economics*, 02 2002, *117* (1), 339–376.

Brown, Jason L., Patrick R. Martin, Donald V. Moser, and Roberto A. Weber, "The Consequences of Hiring Lower-Wage Workers in an Incomplete-Contract Environment," *The Accounting Review*, 2015, *90* (3), 941–966.

Brynjolfsson, Erik and Paul Milgrom, "Complementarities in organizations," *The handbook of organizational economics*, 2013, pp. 11–55.

_ **, Daniel Rock, and Chad Syverson**, "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics," in Ajay Agrawal, Joshua Gans, and Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 2019.

_ **,** _ **, and** _ , "The productivity J-curve: How intangibles complement general purpose technologies.," *American Economic Journal: Macroeconomics*, 2021, *13* (1), 333–72.

_ , **Tom Mitchell, and Daniel Rock**, "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?," *AEA Papers and Proceedings*, May 2018, *108*, 43–47.

Camerer, Colin F. and Roberto Weber, "Growing Organizational Culture in the Laboratory," in Charles R. Plott and Vernon L. Smith, eds., *Handbook of Experimental Economics Results*, Vol. 1 of *Handbook of Experimental Economics Results*, Elsevier, March 2008, chapter 96, pp. 903–907.

_ **and** _ , "Experimental Organizational Economics," in Robert Gibbons and John Roberts, eds., *The Handbook of Organizational Economics*, Princeton University Press, 2012.

Cohen, Michael D and Paul Bacdayan, "Organizational Routines Are Stored As Procedural Memory: Evidence from a Laboratory Study," *Organization Science*, 1994, *5* (4), 554–568.

Dell'Acqua, Fabrizio, "Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters," *Working paper*, 2022.

Deming, David J., "The Growing Importance of Social Skills in the Labor Market," *The Quarterly Journal of Economics*, 2017, *132* (4), 1593–1640.

Dietvorst, B. J., J. P. Simmons, and C. Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of Experimental Psychology: General*, 2015, *144* (1), 114–126.

Englmaier, Florian, Stefan Grimm, David Schindler, and Simeon Schudy, "The Effect of Incentives in Non-Routine Analytical Team Tasks - Evidence from a Field Experiment," *Working paper*, 2018.

Falk, Armin and Andrea Ichino, "Clean Evidence on Peer Effects," *Journal of Labor Economics*, 2006, *24* (1), 39–57.

Felten, Edward W., Manav Raj, and Robert Seamans, "A Method to Link

Advances in Artificial Intelligence to Occupational Abilities," *AEA Papers and Proceedings*, May 2018, *108*, 54–57.

Frey, Carl Benedikt and Michael A. Osborne, "The future of employment: How susceptible are jobs to computerisation?," *Technological Forecasting and Social Change*, 2017, *114*, 254–280.

Gerber, Alan S. and Donald P. Green, *Field Experiments: Design, Analysis, and Interpretation*, W.W. Norton Company, 2012.

Gersick, C J and J R Hackman, "Habitual routines in task-performing groups.," *Organizational behavior and human decision processes*, 1990, *47*, 65–97.

Goldin, Claudia and Lawrence F. Katz, "The origins of technology-skill complementarity," *Quarterly Journal of Economics*, 1998, *3* (4), 693–732.

Goos, Maarten, Alan Manning, and Anna Salomons, "Explaining job polarization: Routine-biased technological change and offshoring," *American Economic Review*, 2014, *104* (8), 2509–26.

Graetz, Georg and Guy Michaels, "Robots at Work," *The Review of Economics and Statistics*, 12 2018, *100* (5), 753–768.

Grip, Andries De and Jan Sauermann, "The Effects of Training on Own and Co-worker Productivity: Evidence from a Field Experiment*," *The Economic Journal*, 2012, *122* (560), 376–399.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis, *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*, OUP Oxford, 2004.

_ , _ , _ , _ , _ , _ , and Richard McElreath, "In search of homo economicus: behavioral experiments in 15 small-scale societies," *American Economic Review*, 2001, *91* (2), 73–78.

Hsieh, Te-Yi, Bishakha Chaudhury, and Emily S Cross, "Human-robot

cooperation in economic games: People show strong reciprocity but conditional prosociality toward robots," 2020.

Jackson, C. Kirabo and Elias Bruegmann, "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers," *American Economic Journal: Applied Economics*, October 2009, *1* (4), 85–108.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshira, *Introductionn to Statistical Learning*, Springer Texts in Statistics, 2013.

Järvelä, Simon, Inger Ekman, J Matias Kivikangas, and Niklas Ravaja, "A practical guide to using digital games as an experiment stimulus," *Transactions of the Digital Games Research Association*, 2014, *1* (2).

Ju, Wendy, "The design of implicit interactions," *Synthesis Lectures on Human-Centered Informatics*, 2015, *8* (2), 1–93.

Kogut, Bruce and Udo Zander, "Knowledge of the firm, combinative capabilities, and the replication of technology," *Organization science*, 1992, *3* (3), 383–397.

Luo, Xueming, Siliang Tong, Zheng Fang, and Zhe Qu, "Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases," *Marketing Science*, 2019, *38* (6), 937–947.

March, James G. and Herbert A. Simon, *An Evolutionary Theory of Economics Change*, New York: John Wiley and Sons, Inc., 1958.

Mas, Alexandre and Enrico Moretti, "Peers at work," *American Economic Review*, 2009, *1* (99), 112–45.

Meier, Stephan, Matthew Stephenson, and Patryk Perkowski, "Culture of trust and division of labor in nonhierarchical teams," *Strategic Management Journal*, 2019, *40* (8), 1171–1193.

Milgrom, Paul and John Roberts, "The Economics of Modern Manufacturing: Technology, Strategy, and Organization," *American Economic Review*, 1990, *80* (3).

Nelson, Richard and Sidney Winter, *An Evolutionary Theory of Economics Change* 1982.

Pope, Devin G. and Justin R. Sydnor, "Implementing anti-discrimination policies in statistical profiling models," *American Economic Journal: Economic Policy*, 2011, *3* (3), 206–231.

Raj, Manav and Robert Seamans, "Primer on artificial intelligence and robotics," *Journal of Organization Design*, 2019, *8* (1), 1–14.

Rick, Scott, Roberto A. Weber, and Colin F. Camerer, "Knowledge Transfer in Simple Laboratory Firms: The Role of Tacit vs. Explicit Knowledge," *Working paper*, 2007.

Trist, Eric Lansdown and Ken W Bamforth, "Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system," *Human relations*, 1951, *4* (1), 3–38.

Washburn, David A, "The games psychologists play (and the data they provide)," *Behavior Research Methods, Instruments, & Computers*, 2003, *35* (2), 185–193.

Webb, Michael, "The Impact of Artificial Intelligence onn the Labor Market," *Working Paper*, 2020.

Weber, Roberto A. and Colin F. Camerer, "Cultural conflict and merger failure: An experimental approach," *Management Science*, 2003, *49* (4), 400–415.

Weidmann, Ben and David J. Deming., "Team Players: How Social Skills Improve Group Performance," *National Bureau of Economic Research*, 2020, *w27071.*

Yannakakis, Georgios N. and Julian Togelius, *Artificial Intelligence and Games*, 1 ed., Springer International Publishing, 2018.

Tables

Table 1: **Descriptive Statistics**

Panel A: Participant-level

	Mean	Std.Dev	Min	P25	Median	P75	Max	N
Skills								
Game 1	38.30	21.94	13.52	23.48	30.73	45.44	99.99	220
Game 2	37.44	10.59	25.72	30.65	34.31	40.31	99.99	220
Game 3	16.15	1.95	13.35	14.79	15.63	17.00	24.90	220
Game 4	18.37	9.67	0.00	11.50	18.00	25.00	51.00	220

Panel B: Team-level

	Mean	Std.Dev	Min	P25	Median	P75	Max	N
Performance outcomes								
Ingredients returned	12.41	2.75	3.00	11.00	12.00	14.00	21.00	1320
Recipes completed	3.83	0.97	1.00	3.00	4.00	4.00	7.00	1320
Coordination failures	6.66	4.00	0.00	4.00	6.00	10.00	26.00	1272
Other variables								
Round difficulty, z-score	-0.00	1.00	-2.12	-0.66	0.05	0.60	2.07	1320
Team skill, z-score	0.03	0.68	-2.11	-0.41	0.13	0.53	1.29	1320

Notes: This table displays descriptive statistics for our sample. Panel A displays descriptive statistics at the participant level (from Phase 1 of the experiment), while Panel B displays them at the team-level (from Phases 2 and 3 of the experiment).

Table 2: **Effect of AI and New hires on ingredients returned**

	DV = Total ingredients returned							
	Total effect				Direct and spillover effect			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI	-0.51***	-1.01**	-0.52**	-0.14				
	(0.19)	(0.44)	(0.20)	(0.43)				
New hire	-0.74***	-1.35***	-0.78***	-0.077				
	(0.25)	(0.44)	(0.27)	(0.42)				
AI direct team					-0.80***	-1.02*	-0.84***	-0.55
					(0.24)	(0.52)	(0.22)	(0.51)
AI spillover team					-0.31	-1.14**	-0.29	0.17
					(0.21)	(0.52)	(0.23)	(0.52)
New hire direct team					-1.29***	-2.52***	-1.25***	-0.35
					(0.34)	(0.53)	(0.35)	(0.58)
New hire spillover team					-0.24	-0.26	-0.35	0.16
					(0.31)	(0.56)	(0.31)	(0.47)
R2	0.469	0.506	0.461	0.530	0.485	0.557	0.475	0.541
Observations	660	110	440	110	660	110	440	110
Control mean	12.992	13.300	12.887	13.100	12.992	13.300	12.887	13.100
Rounds	Phase 3	7	8-11	12	Phase 3	7	8-11	12
P-values, AI = New Hire								
Overall	0.288	0.451	0.304	0.893				
Direct					0.115	0.012	0.219	0.724
Spillover					0.808	0.186	0.836	0.983

Notes: This table examines the effect of automation and new hires on team performance. Columns 1–4 display the results of Equation 2, while Columns 5–8 display the results of Equation 3. Each column refers to a different time period of the task. Columns 1 and 5 display the effect across all six rounds in Phase 3, columns 2 and 6 limit the sample to only round 7 (the round following the manipulation), columns 3 and 7 display the effect for rounds 8–11, and columns 4 and 8 limit the sample to the final round of game play. The bottom of the table displays p-values from an F-test comparing the effects of automation versus the effects of a new hire. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 3: **Effect of AI and new hires on coordination failures**

	DV = # of coordination failures							
	Total effect				Direct and spillover effect			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI	1.28**	2.51**	1.43**	-0.29				
	(0.60)	(1.01)	(0.58)	(0.90)				
New hire	-0.19	-0.92	-0.19	0.64				
	(0.58)	(1.02)	(0.57)	(0.91)				
AI direct team					2.81***	3.80***	3.06***	0.93
					(0.61)	(1.31)	(0.61)	(1.23)
AI spillover team					0.0081	1.33	0.070	-1.22
					(0.68)	(1.10)	(0.70)	(0.94)
New hire direct team					0.48	-0.16	0.49	1.20
					(0.68)	(1.24)	(0.69)	(1.14)
New hire spillover team					-0.77	-1.63	-0.75	0.14
					(0.62)	(1.09)	(0.58)	(1.00)
R2	0.065	0.136	0.070	0.062	0.110	0.174	0.120	0.089
Observations	636	106	424	106	636	106	424	106
Control mean	6.117	5.900	6.075	6.500	6.117	5.900	6.075	6.500
Rounds	Phase 3	7	8-11	12	Phase 3	7	8-11	12
P-values, AI = New Hire								
Overall	0.015	0.001	0.007	0.394				
Direct					0.001	0.007	0.001	0.858
Spillover					0.294	0.012	0.278	0.220

Notes: This table examines the effect of automation and new hires on coordination failures. Columns 1–4 display the results of Equation 2, while Columns 5–8 display the results of Equation 3. Each column refers to a different time period of the task. Columns 1 and 5 display the effect across all six rounds in Phase 3, columns 2 and 6 limit the sample to only round 7 (the round following the manipulation), columns 3 and 7 display the effect for rounds 8–11, and columns 4 and 8 limit the sample to the final round of game play. The bottom of the table displays p-values from an F-test comparing the effects of automation versus the effects of a new hire. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 4: **Impact on trust, effort, and AI attitudes**

	Trust index	Effort index	AI attitudes index
	(1)	(2)	(3)
AI partner	-1.44***	-0.81**	-0.069
	(0.42)	(0.32)	(0.28)
AI spillover	-0.15	-0.24	-0.15
	(0.14)	(0.18)	(0.19)
New hire player	0.028	-0.23	-0.75**
	(0.21)	(0.38)	(0.28)
New hire partner	0.039	-0.26	-0.27
	(0.19)	(0.27)	(0.22)
New hire spillover	0.073	-0.22	-0.094
	(0.17)	(0.24)	(0.21)
R2	0.229	0.056	0.057
Observations	200	200	200

Notes: This table examines the effect of automation and new hires on survey measures of trust, effort, and AI attitudes. Each column displays the results of Equation 3 with $NewHire_{t,f}^{Direct}$ broken down in the new hire and the remaining player whose partner is the new hire. The regression includes controls for round difficulty, player skills, and player performance in phase 3. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 5: **Skills analysis**

Panel A: Heterogeneity by skill

	DV = Total ingredients returned					
	Player replaced			Players remaining		
	Low	Middle	High	Low	Middle	High
	(1)	(2)	(3)	(4)	(5)	(6)
AI	-0.702	-0.242	-0.590**	-0.500*	-0.626**	0.272
	(0.414)	(0.473)	(0.223)	(0.244)	(0.265)	(0.254)
New Hire	-0.993*	-0.363	-1.028***	-0.994***	-1.495***	0.541
	(0.563)	(0.542)	(0.283)	(0.334)	(0.417)	(0.314)
R2	0.461	0.424	0.572	0.517	0.473	0.520
Observations	216	228	216	216	216	228
P-values (Low=Middle=High):						
AI		.724			.019	
New Hire		.514			.001	

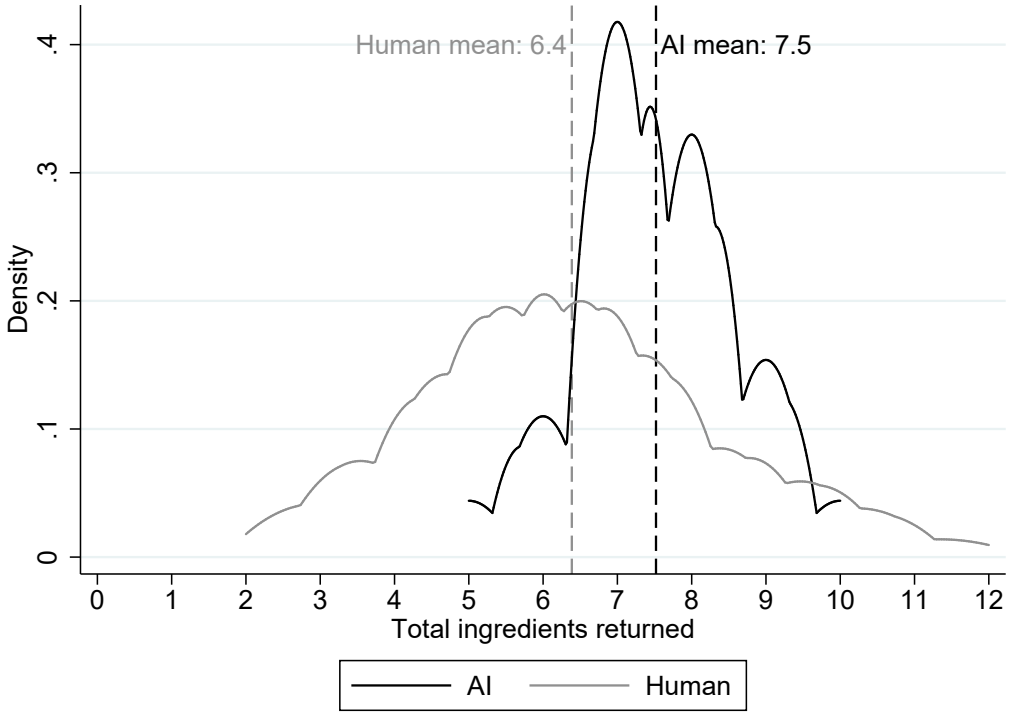
Panel B: Impact of skills on team performance

	Total ingredients	Total ingredients
Player skill, max	0.66*	0.93**
	(0.34)	(0.40)
Player skill, min	0.44***	0.49**
	(0.15)	(0.22)
R2	0.104	0.369
Observations	660	660
Firm fixed effect	No	Yes

Notes: This table examines our results by skills. Panel A examines heterogeneity in the effects of automation and new hires by player skills. We estimate Equation 2 on various sub-samples based on player skills. Low refers to the bottom 1/3, middle to the middle 1/3, and high to the top 1/3 of teams given each skill distribution. Columns 1–3 examine heterogeneity by the skill distribution of the replaced player while columns 4–6 examine heterogeneity by the average skill distribution of the remaining players. The bottom of the table displays p-values from a joint test of treatment effect equality for low, middle, and high-skilled teams. Panel B examines whether the skill of the lowest or highest player is most predictive of team performance. It displays a regression of team output on variables measuring the teams lowest and highest skills, round fixed effects. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

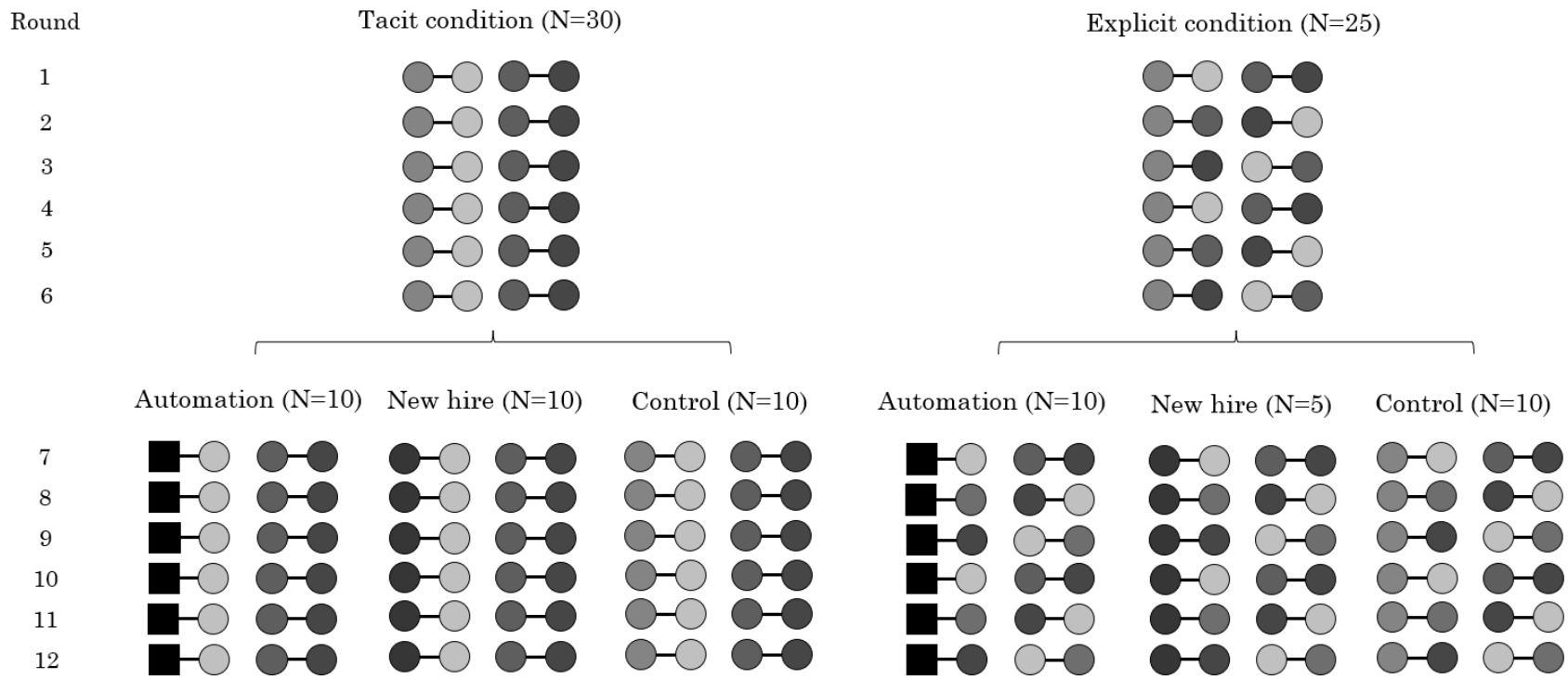
Figures

Figure 1: **Kernel density plot of performance: AI vs humans**



Notes: This figure displays a kernel density plot of performance for human versus automated agents. The data on AI comes from 50 simulations of gameplay without any humans, which lets us observe how the AI would perform in isolation from humans. Meanwhile, the data on human performance comes from round 6 of our experiment, the last round of human play before automated agents are introduced. We exclude the first five rounds (where the performance differentials are even larger) to ensure we are capturing human performance and not game inexperience.

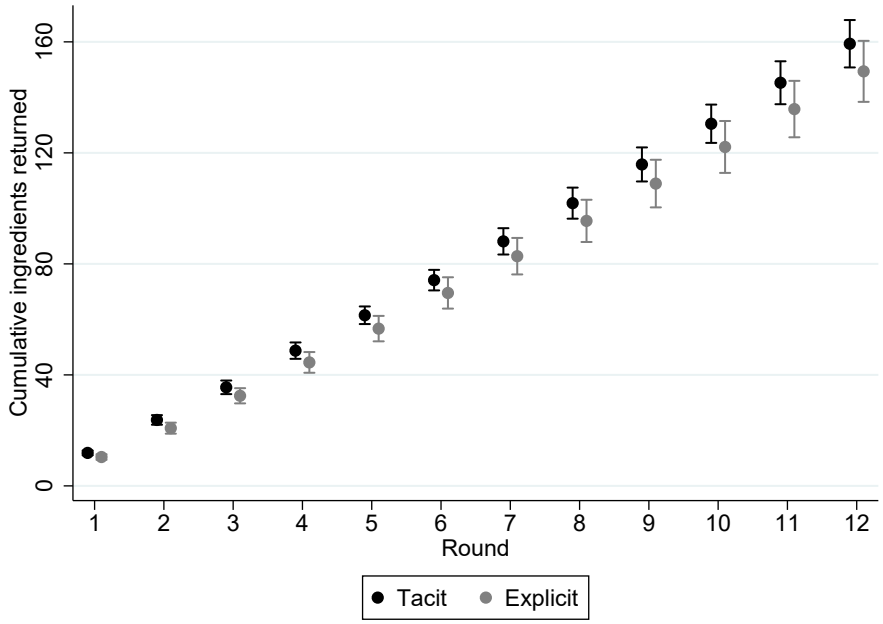
Figure 2: **Experimental Manipulations**



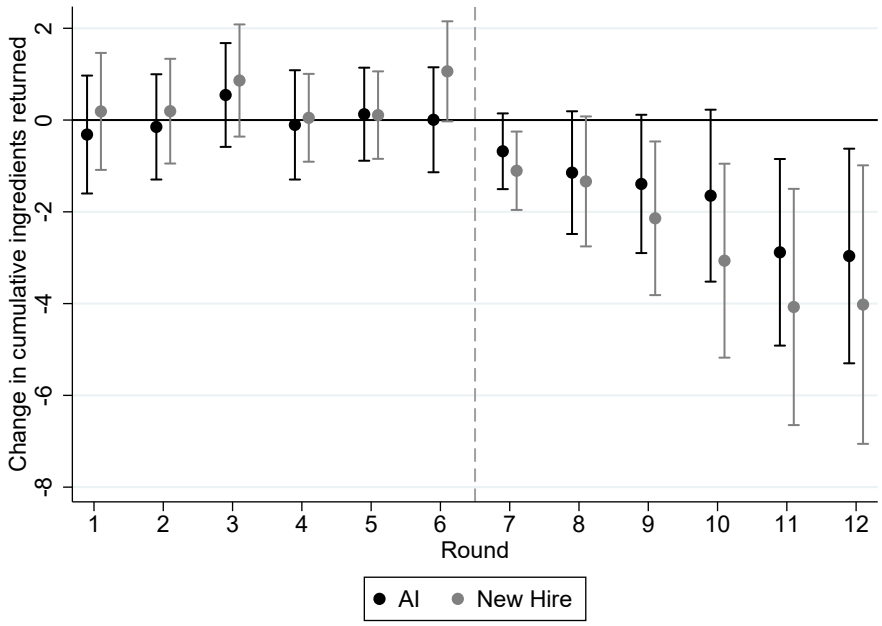
Notes: This figure outlines the experimental conditions and round structure.

Figure 3: **Cumulative performance plots**

(a) Tacit vs explicit, in the control condition



(b) AI and new hire vs control



Notes: This figure displays cumulative ingredients by condition. Figure 5a compares performance in tacit vs explicit teams in the control group. The data comes from a regression of performance on round and tacit/explicit interactions, and phase two controls. Figure 5a plots the cumulative ingredients by round by summing all of the rounds up to and including that round. Figure 5b compares the difference in performance between automation and new hire teams (relative to the control). The data come from two regressions of performance on interactions between round and automation/new hire indicators, using the same linear combination strategy as in Figure 5a.

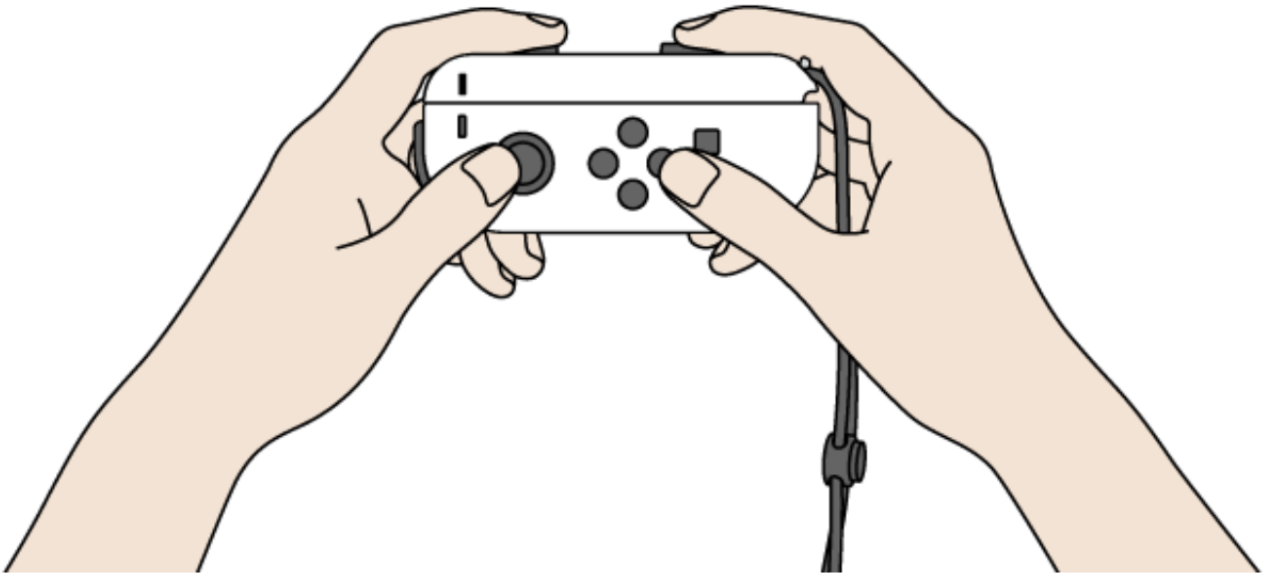
Appendix: For Online Publication Only

A Experimental details

A.1 Nintendo Switch controllers

In order to participate in the game, each participant received a controller like the one displayed in Figure A.1. The controller required two movements for participants. First, participants were required to toggle the joystick using their left thumb to move their character around the map. Second, participants were required to press the right button using their right thumb to pick up and drop off items. An informal survey of participants revealed that not many had prior experience with the game, though all were able to learn the controls easily following phase 1 of the experiment.

Figure A.1: **Nintendo Switch Controller**



Notes: Image taken from https://en-americas-support.nintendo.com/app/answers/detail/a_id/22740/~/how-to-hold-the-joy-con-%28single-and-multiplayer%29

A.2 Survey items following phase 2

The following are the survey items asked following phase 2 of the experiment. Participants responded using a 7-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree."

Please answer truthfully. Your responses will not affect your pay. These questions refer to phase 2 (the previous six rounds). Your partner is the player with whom you played in round 6.

- *Talking with my partner was very important in coordinating our actions.*
- *It is easy to write down a precise set of rules that my partner and I used to coordinate our movement.*
- *My partner and I did not have an explicit plan to determine who was responsible for each recipe ingredient.*
- *I believe that my partner had the appropriate skills and ability to perform well in the game.*
- *Irrespective of their skill level, I believe that my partner exerted maximum effort during the game.*
- *I trusted that my partner performed to the best of their ability during the game.*
- *Irrespective of my skill level, I exerted maximum effort during the game.*
- *I did not pay a lot of attention during the game.*
- *I really tried to perform my best during the game.*

A.3 Survey items following phase 3

The following are the survey items asked following phase 2 of the experiment. Participants responded using a 7-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree."

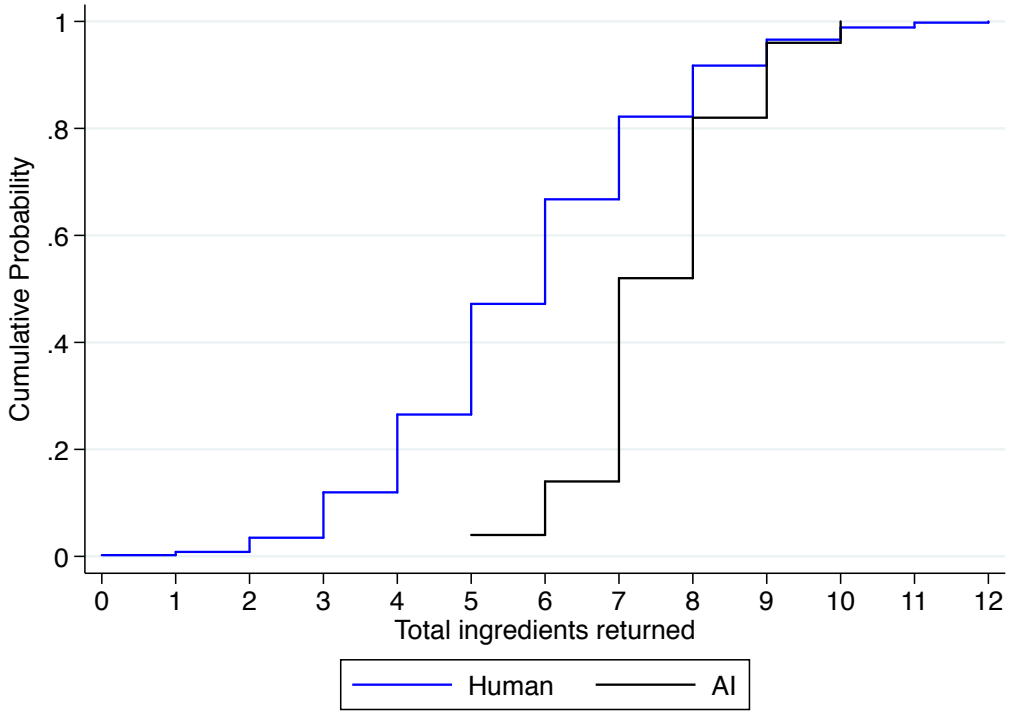
Please answer truthfully. Your responses will not affect your pay. These questions refer to phase 3 (the previous six rounds). Your partner is the player with whom you played in round 12.

- *I believe that my partner had the appropriate skills and ability to perform well in the game.*
- *Irrespective of their skill level, I believe that my partner exerted maximum effort during the game.*
- *I trusted that my partner performed to the best of their ability during the game.*
- *Irrespective of my skill level, I exerted maximum effort during the game.*
- *I did not pay a lot of attention during the game.*
- *I really tried to perform my best during the game.*
- *I believe that artificial intelligence (AI) will have a net positive impact on the world.*
- *I would prefer to perform this task with another human than an automated agent.*
- *I would not feel comfortable taking a ride in an autonomous vehicle.*

A.4 Distribution of performance for human versus automated players

In Figure A.2, we display a cumulative distribution function plot of performance for human versus automated agents. The data on the AI performance comes from 50 simulations of gameplay, where we are able to observe how the AI would perform in isolation from humans. Meanwhile, the data on human performance comes from round 6 of our experiment, the last round of human play before automated agents are introduced. We exclude the first five rounds (where the performance differentials are even larger) to ensure we are capturing human performance and not game inexperience.

Figure A.2: **Distribution of performance**



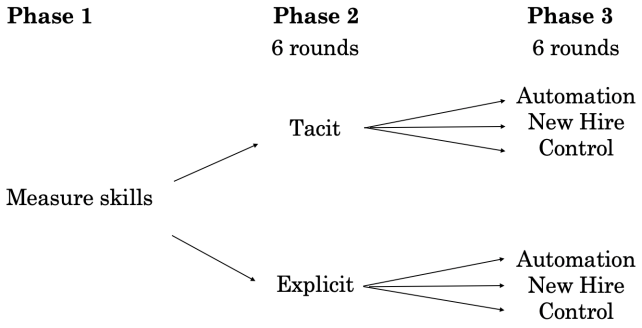
Notes: This figure displays a cumulative distribution function plot of performance for human versus automated agents. The data on the AI performance comes from 50 simulations of gameplay, where we are able to observe how the AI would perform in isolation from humans. Meanwhile, the data on human performance comes from round 6 of our experiment, the last round of human play before automated agents are introduced. We exclude the first five rounds (where the performance differentials are even larger) to ensure we are capturing human performance and not game inexperience.

A.5 Additional Figures

Figure A.3: **Experimental Task**

Notes: This figure displays a screenshot of the experimental task from actual game-play.

Figure A.4: **Structure of Experiment**



Notes: This figure displays the general structure of the experiment. In the first phase of the experiment, each player completed four mini-games from the Super Mario Party game. Prior to the start of the second phase, we randomized participants into one of the two team types. In some firms, participants played with the same partner for all rounds (tacit), while in others they alternated partners (explicit). Prior to the start of the third phase, we randomized teams into treatments by one of the three organizational changes: Automation, New Hire, and a control.

A.6 Pre-registration

In this section, we display text from our pre-registration with our hypotheses, and document a few departures from the pre-analysis plan.

A.6.1 Hypotheses from the pre-registration

Our four primary hypotheses centered around how tacit versus explicit knowledge affect coordination and performance, and how the introduction of AI onto a team impacted existing organizational routines:

- H1: Automation interferes with the team's existing routines and leads to an immediate drop in performance.
- H2: Teams with more tacit knowledge experience a larger drop in performance than teams with more explicit knowledge.
- H3: In the long-run, team performance increases as players learn to coordinate with the automated agent.

- H4: Whether teams perform better with the AI or other humans depends on the skill make-up of the players.

Our secondary hypothesis aimed to test whether the phenomena observed in H1–H3 are unique to AI or would also occur when a new player joins the team.

- SH1: The introduction of AI onto a team is not the same as introducing a human player onto a team.

We aimed to test SH1 by (i) comparing the spillover effects of AI vs human agents, (ii) comparing the direct effects for AI vs human players across the tacit/explicit condition, and (iii) comparing the direct effects for AI vs human players across the AI difficulty conditions.

A.6.2 Pre-analysis departures

In the final version of the manuscript, we decided to focus mainly on three of our four primary pre-registration hypotheses, plus the secondary hypothesis. However, we display the results of all hypotheses.

- (H1 and H3): What are the effects of automation on team performance? Given that team-work requires coordination amongst members, we are also interested in how automation impacts coordination failures. We test H1 and H3 in our results regarding the impacts of automation on team performance and coordination failures, both overall and across various rounds.

- (H2) Do teams in the tacit condition report a larger decrease in performance following automation, relative to teams in the explicit condition? We test for H2 and display those results in the manuscript, but we do not focus on it in our main results. As reported in the main text, our results indicated no difference

in performance for tacit versus explicit teams following Phase 2 (H2 in the pre-registration). Fully developing routines in a team before we randomly treated some teams with an automated partner proved to be very difficult.[1] However, the results in Table B.3 of the main text are in line with H2 in the pre-registration: they show that tacit AI teams experience a larger drop in performance than explicit AI teams, though this difference is not statistically significant.

- (H4): How does the impact of automation vary by the skills of the players and of the team? We test H4 in our results regarding skills.

- Does automation impact the motivation and beliefs of the remaining players? Our results show that automation negatively affects effort and trust in human players. This was not among our main hypotheses, but the pre-registration included measuring trust and effort with a survey. Thus, we had the possibility of staying within our original design and testing for human behavioral responses.

Another departure from the pre-analysis plan is that we decided to pool subjects in the tacit and explicit conditions. We did this for three reasons (described below), although our manuscript displays the results for tacit versus explicit teams separately in Section B.3.

- Results from Phase 2 indicate that the tacit manipulation did not impact team performance or use of explicit communication in Phase 2 (see Appendix Section

[1] Our players played the game for one hour in total. In our pre-testing, we had players play 18 rounds in total, but this was too long: we then reduced it to 12 rounds in the final design, to avoid players growing too tired by the end of the game. This meant that Phase 2 of the design (which was meant to develop routines in a subset of teams) lasted only 6 rounds. As it appears in our results, this proved insufficient to fully develop routines and elicit any difference in performance between the tacit and explicit groups after round 6.

B.1).

- Results from Phase 3 indicate that the effects of automation and turnover do not vary by team type (see Section B.3).

- Columbia University ' s Institutional Review Board closed all non-essential in-person lab studies due to COVID-19 precautions, forcing us to halt data collection on March 12, 2020. At this point, we had finished collecting data on 55/60 teams. The remaining groups were all in the explicit - new hire condition, so the estimates from this condition would be under-powered.

B Additional empirical results

Each of the subsections below goes into greater depth into one of our analyses, or provides an alternative narrative that we analyze. In particular, we focus on the effects of organizational type manipulation, performance plots for direct and spillover teams, validation and analysis of our measure of coordination failures, analysis of survey questions, and a variety of player skills analyses.

B.1 Effect of team type manipulation

In this subsection, we explore the impact of our team type manipulation. We first examine how team type influenced team performance. Teams in the tacit condition played with the same partner for all 12 rounds, while those in the explicit condition alternated partners every round. Table B.1 displays the impact of this manipulation on team performance across all 12 rounds (columns 1 and 4), in phase 2 (rounds 1–6; columns 2 and 5), and in phase 3 (rounds 7–12; columns 3 and 6). Columns 1–3 pool across the organizational change conditions (i.e., control for assignment to the AI and new hire conditions) while columns 4–6 limit the same only to control units who did not receive an AI or a new hire.

The results show that teams in the tacit condition returned slightly more ingredients per round, though this result is not statistically significant at conventional levels. The differences in performance are weakly significant in phase 3, the final six rounds of play.

Table B.1: **Effect of tacit manipulation on team performance**

	DV = Total ingredients					
	All firms			Control firms only		
	(1)	(2)	(3)	(4)	(5)	(6)
Tacit (same partner)	0.245	0.0451	0.325*	0.743	0.610	0.482*
	(0.315)	(0.350)	(0.174)	(0.475)	(0.484)	(0.257)
R2	0.341	0.326	0.454	0.374	0.311	0.402
Observations	1320	660	660	480	240	240
Control mean	12.617	11.475	13.758	12.617	11.475	13.758
Rounds	1-12	Phase 2	Phase 3	1-12	Phase 2	Phase 3

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.010

Notes: This table examines the effect of the tacit manipulation on team performance using equation 1. The table displays the impact of this manipulation on team performance across all 12 rounds (columns 1 and 4), in phase 2 (rounds 1–6; columns 2 and 5), and in phase 3 (rounds 7–12; columns 3 and 6). Columns 1–3 pool across the organizational change conditions (i.e., control for assignment to the AI and new hire conditions) while columns 4–6 limit the same only to control units who did not receive an AI or a new hire. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

We additionally explore the impact of our team type manipulation on survey questions related to explicit communication. We estimate equation 1 and display the results in Table B.2. Column 1 displays the results on an aggregate z-score of explicit communication, while columns 2–4 show the impact on individual survey items. The results indicate that teams in the tacit condition did not score differently on our explicit communication index (column 1, $p = 0.38$). We find evidence that the manipulation impacted how important verbal communication was for coordinating behavior (column 2), but there was no impact on using precise rules or explicit plans for coordinating movement or gathering ingredients (columns 3 and 4).

Table B.2: **Effect of tacit manipulation on explicit communication**

	Use of explicit communication z-score	Talking was important in coordinating actions	Precise rules to coordinate movement	Explicit plan for ingredients
Tacit condition	-0.14	-0.42**	0.30	0.34
	(0.16)	(0.20)	(0.27)	(0.34)
R2	0.009	0.028	0.009	0.011
Observations	220	220	220	220

Notes: This table examines the effect of the tacit manipulation on use of explicit communication using equation 1.Column 1 displays the results on an aggregate z-score of explicit communication, while columns 2–4 show the impact on individual survey items. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

B.2 Aggregate performance plots for direct and spillover teams

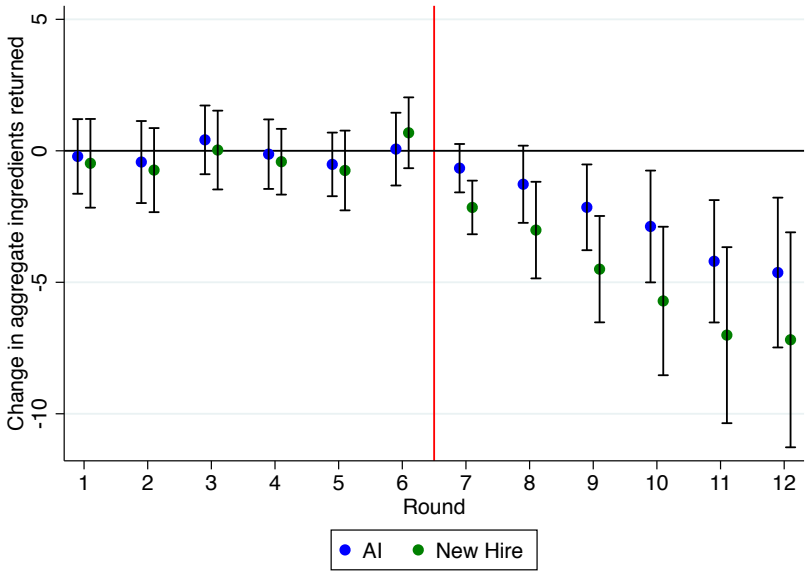
In Figure 3b of the main text, we display aggregate performance plots for the overall effects of the automation and new hire conditions. In Figures B.5 and B.6, we display them for the direct and spillover effects, respectively.

B.3 Heterogeneity by team type and task difficulty

Team type. We first examine heterogeneity by team type. As discussed in Section 2.3, our experimental generation of tacit knowledge is modelled on the Rick et al. (2007) experiment. Accordingly, we randomized participants to either play with the same partner throughout all 12 rounds (tacit condition) or switch partners each round (explicit condition). Column 1 of Table B.3 examines differences in the impact of automation across these teams. We find limited evidence that the effects of automation differed by team type. Although automation has a larger (more negative) effect on tacit teams, this difference is not statistically significant (p=0.55). We also find limited evidence that the effect of new hires differed by team type (p=0.64).

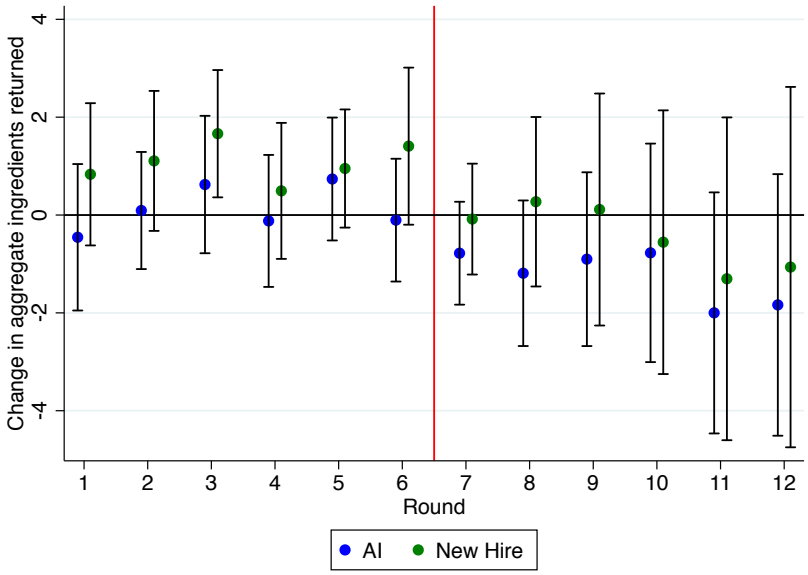
Task difficulty. Second, we examine heterogeneity by the difficulty of the coordination task for which we have developed an innovative measure. Although teams have to

Figure B.5: **Aggregate performance plot, directly-impacted teams**



Notes:This figure displays the aggregate performance plot for teams that receive an automated agent or a new hire relative to control teams. The data come from two regressions of performance on interactions between round and automation/new hire direct effect indicators (one for phase 2 differences, using phase 2 controls, and one for phase 3 differences, using phase 3 controls). The coefficients in phase 3 come from the same linear combination strategy as in Figure 3b.

Figure B.6: **Aggregate performance plot, spillover teams**



Notes: This figure displays the aggregate performance plot for AI and new hire spillover teams relative to control teams. The data come from two regressions of performance on interactions between round and automation/new hire spillover effect indicators (one for phase 2 differences, using phase 2 controls, and one for phase 3 differences, using phase 3 controls). The coefficients in phase 3 come from the same linear combination strategy as in Figure 3b.

complete the same task during each round of gameplay (gathering three ingredients to complete a recipe), the recipes differ in their difficulty. For example, in Figure A.3, it is more difficult for the team on the left to complete a recipe requiring two lettuce and one tomato versus one requiring two bacon and one tomato. There are a total of seven possible recipes, and we can leverage their random assignment to create a measure of task difficulty. We recorded the game screen and had a research assistant code up the recipes that appear in each round.[2] We then create a round difficulty measure by regressing team performance on a set of binary indicators corresponding to each of the seven recipe types plus round fixed effects, firm fixed effects, and the total number of recipes assigned to each team.[3] We limit the regression to Phase 2 performance (prior to the introduction of the automated agent and new hires) and generate predicted values using only the coefficients attached to the seven binary indicators. We use controls for the count of recipes assigned, round fixed effects, and firm fixed effects in the regression but not in the estimation of the round difficulty index. In the prediction we set all non-binary variables to zero so that the prediction only captures the relationship between the recipe make-up, and not differences caused by the skill make-up of the players (the organization fixed effects) or player learning over time (the round fixed effects).[4]

We run two sets of analyses to validate our round difficulty measure. We display these in Appendix B.12 but briefly mention them here. First, we examine the weights

[2]The seven recipes are: (i) bacon-lettuce-tomato ($p = 2/8$), (ii) bacon-bacon-tomato ($p = 1/8$), (iii) bacon-bacon-lettuce ($p = 1/8$), (iv) bacon-tomato-tomato ($p = 1/8$), (v) bacon-lettuce-lettuce ($p = 1/8$), (vi) tomato-tomato-lettuce ($p = 1/8$), and (vii) tomato-lettuce-lettuce ($p = 1/8$).

[3]We control for the total number of recipes assigned to each pair because higher performing teams are assigned more recipes because they return more ingredients. If we do not control for this, each of the binary indicators returns positive coefficients. By controlling for the total number of recipes assigned to each team, the regression returns: conditional on a team being assigned X recipes, how does the make-up of those recipes influence whether or not the team completes the Xth recipe?

[4]See Pope and Sydnor (2011) for more information about this method.

placed on each of the seven types of recipes in the initial prediction. We find that the recipes with the largest negative weights are those which contain two ingredients that are farthest from a team, which lines up with our prior that teams are more likely to return ingredients when they are closest to their table. Second, we validate the measure by testing how well it predicts team performance in a given round. We find that the relationship between the round difficulty measure and team performance is negative and statistically significant, even when we control for round and firm fixed effects. For these two reasons, we believe our index is an accurate measure of task difficulty.

Table B.3 examines heterogeneity by round difficulty in Column 2. The results indicate that the effects of automation do not vary by the difficulty of the task ($p = 0.40$). Meanwhile, the impact of human turnover depends on the difficulty of the task. Introducing a new human player onto a team in a round that is one standard deviation higher in round difficulty further decreases team performance by 0.26 ingredients ($p = 0.07$) Although automation decreases performance in our setting, we observe that the introduction of an automated agent decreases variance in performance and flattens it, especially compared to the introduction of new human players who are more susceptible to task difficulty.

B.4 Validation of our coordination failure measure

In the main text of the paper, we argue that coordination failures hamper firm performance. Our measure of coordination failures is the number of bumps between team members. We validate our coordination failure measure in Table B.4, where we regress team performance on coordination failures. Column 1 contains no controls, column 2 controls for round fixed effects, column 3 controls for round fixed effects and treatment identifiers, and column 4 controls for round fixed effects and firm fixed effects. The results indicate that teams who have more coordination failures return fewer

Table B.3: **Heterogeneity by team structure and task difficulty**

	Total ingredients	Total ingredients
AI	-0.40**	-0.53***
	(0.18)	(0.19)
AI X tacit	-0.21	
	(0.36)	
New hire	-0.60	-0.74***
	(0.39)	(0.25)
New hire X tacit	-0.23	
	(0.49)	
AI X round difficulty		0.12
		(0.14)
New hire X round difficulty		-0.26*
		(0.14)
R2	0.470	0.473
Observations	660	660
Control mean	13.375	6.050
rounds	7-12	7-12

Notes: This table examines heterogeneity in the effects of automation and new hires by team structure and round difficulty using Equation 2, which regresses total performance on binary indicators of assignment to AI/New Hire, round fixed effects, round difficulty, and team controls included team skills, phase 2 performance, and a binary indicator for assignment to the tacit condition. We interact the AI and new hire coefficients with the tacit indicator in column 1, and with the round difficulty indicator in column 2. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

ingredients. This is even true when we include fixed effects for each firm: teams within the same firm return fewer ingredients in rounds where they have more coordination failures.

Table B.4: **Correlation between coordination failures and team ingredients returned**

	(1)	(2)	(3)	(4)
	Total ingredients			
Coordination failures	-0.0887***	-0.0532**	-0.0526**	-0.0475**
	(0.0245)	(0.0219)	(0.0229)	(0.0229)
R2	0.016	0.230	0.232	0.423
Observations	1272	1272	1272	1272
Round f.e	No	Yes	Yes	Yes
Treatment	No	No	Yes	No
Team f.e.	No	No	No	Yes

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.010

Notes: This table examines the correlation between coordination failures and total ingredients returned. It displays the results of regressions of ingredients returned on coordination failures with different fixed effects and controls in each column. Column 1 contains no controls, column 2 controls for round fixed effects, column 3 controls for round fixed effects and treatment identifiers, and column 4 controls for round fixed effects and firm fixed effects. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

B.5 Effect of tacit manipulation on survey indices

In Table 4 of the main text, we document that participants assigned to play with the AI reported lower trust and lower effort. We had players respond to this questionnaire based off their partner in the final round of gameplay, and whether players were in the tacit or explicit condition influenced how often they directly collaborated with their partner: those in the tacit condition had six rounds of gameplay in phase 3 with their partner, while those in the explicit group only had two. In theory, it is likely that this influenced their attitudes and beliefs, especially about trust in their partner. However, our results indicate very limited heterogeneity by the team type condition. In Table B.9 below, we reproduce Table 4 from the main text but with a formal test for

heterogenous effects, whereby we interact each participant role with the tacit condition. The role*tacit variables capture whether there are any differential impacts on our survey items by the tacit condition. Our results indicate very limited differences in the impact of automation on trust and effort across the two team types. Regardless of whether they played with the automated agent for two or six rounds, players reported lower trust in their automated partner and lower individual effort.

B.6 Effect of automation on specific survey items

In Table 4 of the main text, we examine treatment effects on our (aggregated) survey indices. In this section, we display the results for individual survey questions. Table B.6 displays the effects for our trust questions, Table B.7 for our effort questions, and Table B.8 for our questions on attitudes toward AI.

B.7 Impact on trust, effort, and AI attitudes, by tacit or explicit team type

In Table 4 of the main text, we examine the effects of automation on our aggregated survey indices. This regression pools the tacit and explicit conditions and returns the average treatment effect on the survey indices across both the tacit and explicit conditions. However, whether players were in the tacit or explicit condition influenced how often they directly collaborated with their partner: those in the tacit condition had six rounds of gameplay in phase 3 with their partner while those in the explicit group only had two. This could influence player responses to the survey items, if these responses are affected by how many times respondents collaborated with their partner. For example, the decrease in trust could be concentrated in the explicit condition, while those in the tacit condition learned to trust the AI after six rounds of gameplay.

In this section, we investigate whether the behavioral and attitudinal responses to automation vary by whether the player was assigned to a tacit or explicit team. We

Table B.5: **Impact on trust, effort, and AI attitudes, by tacit and explicit conditions**

	Trust in partner, z-score	Own effort, z-score	Attitudes toward AI, z-score
AI partner	-1.29**	-0.66*	-0.12
	(0.58)	(0.36)	(0.35)
AI partner X tacit	-0.30	-0.28	0.095
	(0.84)	(0.63)	(0.56)
AI spillover	-0.15	-0.16	-0.17
	(0.17)	(0.21)	(0.30)
AI spillover X tacit	-0.0048	-0.16	0.028
	(0.29)	(0.36)	(0.38)
New hire player	0.19	-0.91	-1.03*
	(0.28)	(0.82)	(0.51)
New hire player X tacit	-0.25	1.02	0.42
	(0.40)	(0.89)	(0.62)
New hire partner	-0.32	-0.73	-0.033
	(0.35)	(0.47)	(0.48)
New hire partner X tacit	0.54	0.71	-0.34
	(0.41)	(0.56)	(0.52)
New hire spillover	-0.045	0.089	-0.34
	(0.31)	(0.23)	(0.30)
New hire spillover X tacit	0.18	-0.47	0.39
	(0.37)	(0.41)	(0.41)
Tacit condition	-0.17	0.082	-0.056
	(0.15)	(0.21)	(0.26)
R2	0.238	0.092	0.066
Observations	200	200	200

Notes: This table examines the effect of automation and new hires on survey measures of trust, effort, and AI attitudes by the tacit and explicit conditions. Each column displays the results of Equation 3 with $NewHire_{t,f}^{Direct}$ broken down in the new hire and the remaining player whose partner is the new hire, and then interacts all the player roles with the tacit indicator. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table B.6: **Effect on trust items**

	Partner had appropriate skills, + agree	Partner exerted appropriate effort, + agree	I trusted my partner, + agree
AI partner	-0.55	-1.37***	-1.23***
	(0.34)	(0.38)	(0.34)
AI spillover	-0.077	-0.062	-0.19
	(0.15)	(0.098)	(0.13)
New hire player	0.17	-0.014	-0.086
	(0.17)	(0.17)	(0.17)
New hire partner	-0.12	0.099	0.10
	(0.22)	(0.14)	(0.13)
New hire spillover	0.15	0.053	-0.033
	(0.14)	(0.12)	(0.16)
R2	0.079	0.277	0.222
Observations	200	200	200

Notes: This table examines the effect of automation and new hires on the three survey questions related to trust. Each column displays the results of Equation 3 with $NewHire_{t,f}^{Direct}$ broken down in the new hire and the remaining player whose partner is the new hire. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table B.7: **Effect on effort items**

	I exerted maximum effort, + agree	I paid no attention, + agree	I tried my best, + agree
AI partner	-0.28*	1.00**	-0.27*
	(0.15)	(0.47)	(0.14)
AI spillover	-0.098	0.27	-0.074
	(0.080)	(0.27)	(0.083)
New hire player	-0.24	0.27	0.058
	(0.25)	(0.42)	(0.11)
New hire partner	-0.032	0.30	-0.15
	(0.12)	(0.42)	(0.14)
New hire spillover	-0.024	0.38	-0.087
	(0.090)	(0.30)	(0.12)
R2	0.039	0.060	0.043
Observations	200	200	200

Notes: This table examines the effect of automation and new hires on the three survey questions related to effort. Each column displays the results of Equation 3 with $NewHire_{t,f}^{Direct}$ broken down in the new hire and the remaining player whose partner is the new hire. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table B.8: **Effect on AI attitude iems**

	AI will have net positive impact, + agree	I prefer to perform task with human over AI, + agree	I would not feel comfo with autonomous vehicle,
AI partner	0.19	-0.042	0.49
	(0.41)	(0.33)	(0.47)
AI spillover	0.16	0.17	0.42
	(0.27)	(0.24)	(0.34)
New hire player	-1.12**	0.38	0.33
	(0.53)	(0.33)	(0.47)
New hire partner	-0.15	0.52**	-0.11
	(0.32)	(0.24)	(0.45)
New hire spillover	-0.044	0.11	0.073
	(0.24)	(0.25)	(0.31)
R2	0.094	0.039	0.027
Observations	200	200	200

Notes: This table examines the effect of automation and new hires on the three survey questions related to AI attitudes. Each column displays the results of Equation 3 with $NewHire_{t,f}^{Direct}$ broken down in the new hire and the remaining player whose partner is the new hire. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

reproduce Table B.9 from the main text but estimate separate regressions for the tacit and explicit conditions. Table B.9 displays these results, with teams in the odd columns and explicit ones in the even columns. Given that we cut our sample size by about half, the precision in treatments is lower than for the pooled version displayed in Table 4.However, the results indicate very limited differences in the impact of automation on trust and effort across the two team types. Regardless of whether they played with the automated agent for two or six rounds, players reported lower trust in their automated partner and that they reported lower effort.

B.8 Correlation between skills and performance

In this section, we document that our skill measures are strongly correlated with performance. We do so by examining (i) whether individual skills are correlated with individual ingredients returned (Table B.10) and (ii) whether average team skill is

Table B.9: **Impact on trust, effort, and AI attitudes, by tacit and explicit conditions**

	Trust in partner, z-score	Own effort, z-score	Attitudes toward AI, z-score
AI partner	-1.29**	-0.66*	-0.12
	(0.58)	(0.36)	(0.35)
AI partner X tacit	-0.30	-0.28	0.095
	(0.84)	(0.63)	(0.56)
AI spillover	-0.15	-0.16	-0.17
	(0.17)	(0.21)	(0.30)
AI spillover X tacit	-0.0048	-0.16	0.028
	(0.29)	(0.36)	(0.38)
New hire player	0.19	-0.91	-1.03*
	(0.28)	(0.82)	(0.51)
New hire player X tacit	-0.25	1.02	0.42
	(0.40)	(0.89)	(0.62)
New hire partner	-0.32	-0.73	-0.033
	(0.35)	(0.47)	(0.48)
New hire partner X tacit	0.54	0.71	-0.34
	(0.41)	(0.56)	(0.52)
New hire spillover	-0.045	0.089	-0.34
	(0.31)	(0.23)	(0.30)
New hire spillover X tacit	0.18	-0.47	0.39
	(0.37)	(0.41)	(0.41)
Tacit condition	-0.17	0.082	-0.056
	(0.15)	(0.21)	(0.26)
R2	0.238	0.092	0.066
Observations	200	200	200

Notes: This table examines the effect of automation and new hires on survey measures of trust, effort, and AI attitudes by the tacit and explicit conditions. Each column displays the results of Equation 3 with $NewHire_{t,f}^{Direct}$ broken down in the new hire and the remaining player whose partner is the new hire, with tacit results in the odd columns and explicit ones in the even ones. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

correlated with team performance (Table B.11). In both tables, Column 1 contains no controls, column 2 controls for round fixed effects, column 3 controls for round fixed effects and treatment identifiers, and column 4 controls for round fixed effects and firm fixed effects. The results indicate that teams who receive a more difficult set of recipes return fewer ingredients. This is even true when we include fixed effects for each organization: column 4 illustrates that a team returns 0.520 fewer ingredients in a round where they receive recipes that are one standard deviation higher than average, compared to that same team in a round with average round difficulty.

Table B.10: **Correlation between individual skill and individual ingredients returned**

	Ingredients			
	(1)	(2)	(3)	(4)
Skills, z-score	0.562***	0.538***	0.556***	0.563***
	(0.0831)	(0.0828)	(0.0819)	(0.0913)
R2	0.084	0.187	0.192	0.273
Observations	2640	2640	2640	2640
Round fixed effects	No	Yes	Yes	Yes
Treatment indicators	No	No	Yes	No
Firm fixed effects	No	No	No	Yes

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.010

Notes: This table examines the correlation between individual skills and individual ingredients returned. It displays the results of regressions of ingredients returned on our skills index with different fixed effects and controls in each column. Column 1 contains no controls, column 2 controls for round fixed effects, column 3 controls for round fixed effects and treatment identifiers, and column 4 controls for round fixed effects and firm fixed effects. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

B.9 Additional analysis on the complementarity between skills and automation

In this subsection, we present additional analysis on the complementarity between skills and automation that we observed in the study.

Table B.11: **Correlation between team skills and team ingredients returned**

	Total ingredients			
	(1)	(2)	(3)	(4)
Team skills, z-score	1.281***	1.171***	1.262***	1.414***
	(0.269)	(0.271)	(0.268)	(0.344)
R2	0.096	0.303	0.314	0.481
Observations	1320	1320	1320	1320
Round fixed effects	No	Yes	Yes	Yes
Treatment indicators	No	No	Yes	No
Firm fixed effects	No	No	No	Yes

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.010

Notes: This table examines the correlation between average team skills and total ingredients returned. It displays the results of regressions of ingredients returned on our skills index with different fixed effects and controls in each column. Column 1 contains no controls, column 2 controls for round fixed effects, column 3 controls for round fixed effects and treatment identifiers, and column 4 controls for round fixed effects and firm fixed effects. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

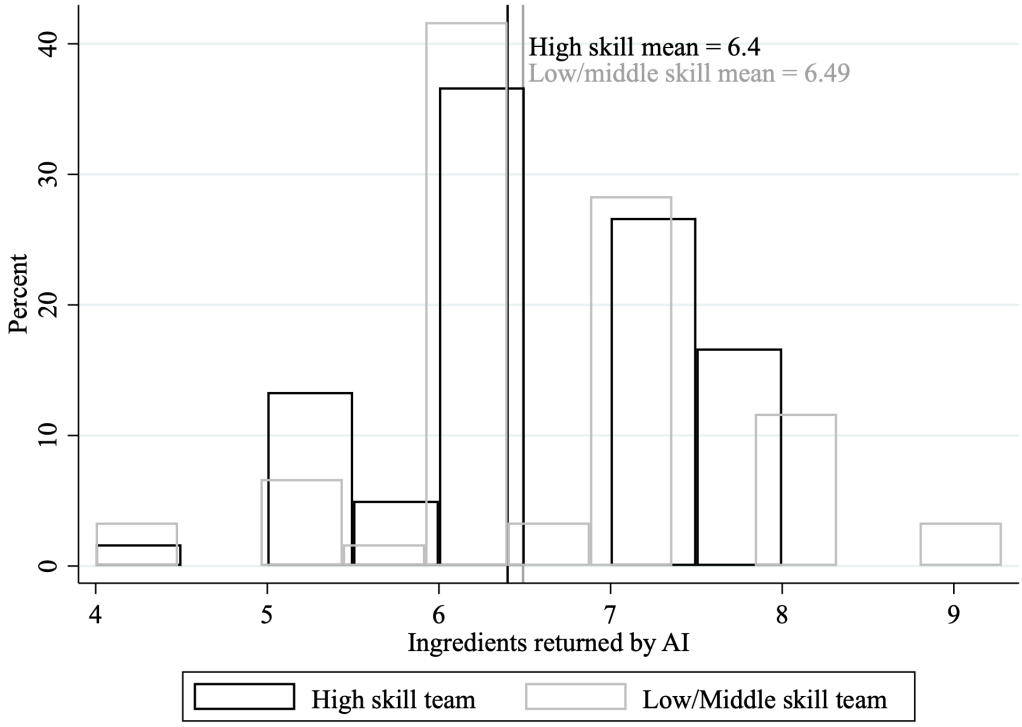
B.9.1 Additional test for complementarities

An alternative test for complementarities is given by Milgrom and Roberts (1990) and Brynjolfsson and Milgrom (2013). Suppose that $y_1 = 1$ if the team is assigned to the automation condition and $y_1 = 0$ if the team is assigned to the control condition, and suppose that $y_2 = 1$ if the team is high-skilled and $y_2 = 0$ if the team is low- or middle-skilled.Finally, let $f(y_1, y_2)$ be the average team performance. Given the exogenous assignment of teams to the automation condition, and of players (and their skills to teams), complementarities exist between y_1 and y_2 if the following equation is satisfied: $f(1,1) - f(0,0) >= f(1,0) - f(0,0) + f(0,1) - f(0,0)$. This can be rewritten as: $f(1,1) >= f(1,0) + f(0,1) - f(0,0)$. If we fill in average performance scores, this condition holds ($13.43 >= 12.18$). In other words, total output is greater when high skilled teams receive the AI, than when either is implemented separately.

B.9.2 Do automated agents perform better on high- vs middle- and low-skilled teams?

We first investigate whether the complementarity arises because automated agents themselves perform better when paired with high-skilled versus low- and medium-skilled partners. This can occur, for example, if the automated agent is able to more easily respond to the actions of high-skilled partners who move more quickly through the map. However, we find no evidence that the complementarity is due to the automated agent returning more ingredients when matched to a high-skilled partner. In Figure B.7, we display a histogram of the number of ingredients returned by the AI when it is matched to a high- versus middle- or low-skilled partner. The results indicate limited differences in the AIs performance (6.4 when paired with high-skilled teams vs 6.5 when paired with middle- and low-skilled teams). We also confirm there are no differences in the AI's performance using a regression in Table B.12. Column 1 displays the results of a regression of AI performance on a binary indicator for being partnered with a high-skilled coworker, while column 2 additionally includes round fixed effects and our round difficulty measure. The results indicate no statistically significant differences in AI performance by the skills of its coworker ($p = 0.61$ and $p = 0.56$, respectively). For this reason, we do not believe the source of complementarity stems from the AI performing better when matched to high-skilled partner.

Figure B.7: **AI performance by team assignment**



Notes:This figure displays a histogram of the number of ingredients returned by the AI when it is paired with a high-skilled versus middle- or low-skilled team in phase 3.

Table B.12: **Impact of team assignment on AI performance**

	(1)	(2)
	Ingredients returned by AI	
High-skilled team	-0.0917	-0.0989
	(0.178)	(0.166)
R2	0.002	0.016
Observations	120	120
Round controls	No	Yes

* p<0.10, ** p<0.05, *** p<0.010

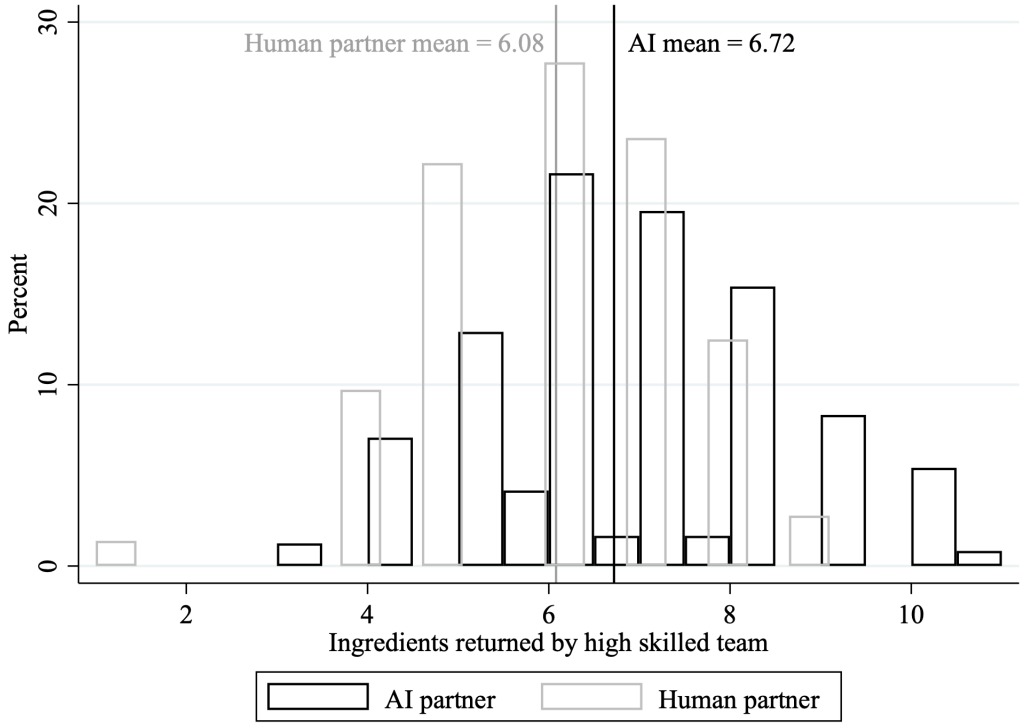
Notes: This table examines the impact of team assignment on AI performance. It compares the number of ingredients returned by the AI when the AI plays alongside a high- versus middle-/low-skilled partner. Column 1 displays the results of a regression of AI performance on a binary indicator for being partnered with a high-skilled coworker, while column 2 additionally includes round fixed effects and our round difficulty measure. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

B.9.3 Do high-skilled firms perform better when they partner with an AI versus another human player?

Another possibility is that high-skilled firms perform better when they are partnered with an automated agent versus another human player. We find evidence in support of this conjecture. In Figure B.8, we display a histogram of the number of ingredients returned by players in high-skilled firms when they are paired with an AI versus a control human player. The results indicate that players in high-skilled firms return more ingredients when paired with an AI (6.7) versus a human player (6.1). This difference is at the player level, so the firm-level difference is 1.8 ingredients per round ($= 0.6 * 3$). We also confirm this using a regression in Table B.13. Column 1 displays the results of a regression of ingredients on a binary indicator for being in the same experimental firm as an AI while column 2 additionally includes round fixed effects and our round difficulty

measure. Both regressions limit the sample to participants in high-skilled firms, so that the coefficient on AI measures the average change in performance for players in high-skilled firms when they are assigned to play with an AI versus a human partner. The results indicate that players on high-skilled firms return on average between 0.70 and 0.80 ingredients more with an AI as their final player versus another human control ($p = 0.01$ and $p = 0.01$, respectively). For that reason, the source of complementarity stems from an improvement in the performance of high-skilled firms when assigned an automated agent versus a human control player.

Figure B.8: **High-skilled teams performance by partner assignment**



Notes:This figure displays a histogram of the number of ingredients returned in phase 3 by players in high-skilled teams when they are paired with an AI versus when they are paired with the same human partner as phase 2.

Table B.13: **Impact of partner assignment on high-skilled team performance**

	(1)	(2)
	Ingredients returned by players on high-skilled teams	
AI	0.633***	0.676***
	(0.202)	(0.205)
R2	0.027	0.045
Observations	312	312
Round controls	No	Yes

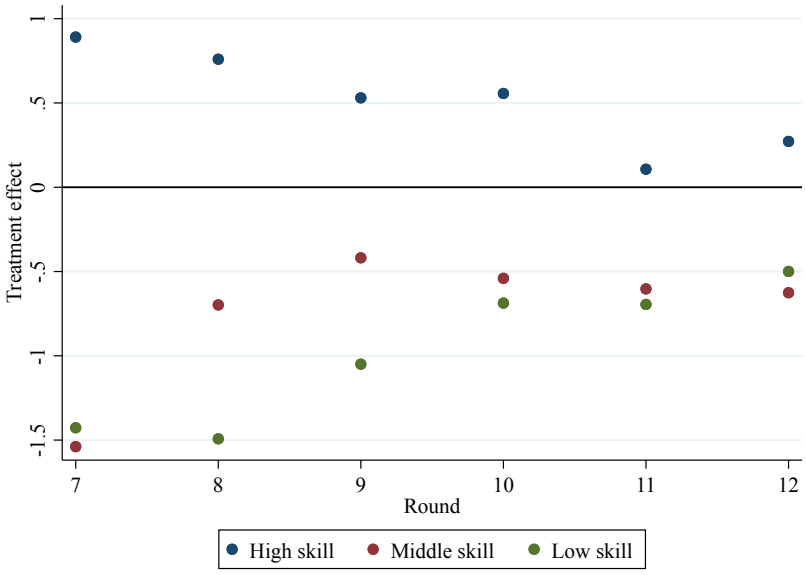
* p<0.10, ** p<0.05, *** p<0.010

Notes: This table examines the impact of partner assignment on the performance of participants in high-skilled firms. It compares the number of ingredients returned by players on a high-skilled firm when they are paired with an AI versus another human control player. Column 1 displays the results of a regression of performance on a binary indicator for being in the same firm as an AI, while column 2 additionally includes round fixed effects and our round difficulty measure. The regression limits the sample to participants in high-skilled firms. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

B.10 Skill heterogeneity by round

In the main text, we document that automation is skill-biased in our experiment; low- and medium-skill firms perform worse following automation whereas high-skill ones do not. In this subsection, we examine these skill effects over time. We run the same regression as in section 4.3 but subset the data by round. Figure B.9 displays the treatment effects by round for each skill category. The results indicate that the difference in treatment effects across the skill distribution occurs immediately upon automation. In fact, the difference in treatment effects is largest in the round following automation. As time goes on, the difference shrinks but it is still statistically significant across all six rounds ($p = 0.02$).

Figure B.9: **Skill heterogeneity by round**



Notes:This figure displays treatment effect across round and by team skill levels.

B.11 The impact of automation on coordination failures, by team skill

In Table B.14, we examine whether partner assignment (AI vs human) affects the number of coordination failures in high-skilled firms. Specifically, we test the hypothesis that high-skilled firms have fewer coordination failures when they receive an automated agent versus when they keep their human player. The table compares the number of coordination failures in high-skilled firms when they are paired with an AI versus another human control player. Column 1 displays the results of a regression of coordination failures on a binary indicator for being partnered with AI, while column 2 additionally includes round fixed effects and our round difficulty measure. Both regressions limit the sample to participants in high-skilled firms, so that the coefficient on AI measures the average change in coordination failures in high-skilled firms when they are assigned to

play with an AI versus a human partner. The results show that partner assignment (AI vs human) has no significant impact on coordination failures. The hypothesis above is thus rejected.

Table B.14: **Impact of partner assignment on coordination failures in high-skilled teams**

	(1)	(2)
	Coordination failures	
AI	1.244	1.310
	(1.197)	(1.206)
R2	0.021	0.036
Observations	156	156
Round controls	No	Yes

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.010

Notes: This table examines the impact of partner assignment on the number of coordination failures in high-skilled firms. It compares the number of coordination failures on a high-skilled firm when they are paired with an AI versus another human control player. Column 1 displays the results of a regression of AI performance on a binary indicator for being partnered with a high-skilled firm, while column 2 additionally includes round fixed effects and our round difficulty measure. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

B.11.1 Heterogeneity in direct and spillover effects by team skill

We also investigate heterogeneity in the direct and spillover effects of automation by team skills. We estimate Equation 3 for high-, middle-, and low-skilled teams in Table B.15. Columns 1–3 examine heterogeneity by the average skill distribution of the remaining players. The bottom of the table displays p-values from a joint test of treatment effect equality for low, middle, and high-skilled teams (for both the direct and spillover effects). Although the estimates are noisier, the results indicate that the

high-skilled teams are better able to absorb both the direct and spillover effects of automation relative to low- and middle-skilled teams This indicates that both channels matter in explaining the superior performance of high-skilled teams.

Table B.15: **Heterogeneity in direct and spillover effects by team**

	Total Ingredients		
	(1)	(2)	(3)
AI direct team	-0.853**	-0.728*	-0.007
	(0.314)	(0.378)	(0.295)
AI spillover team	-0.433	-0.588**	0.571**
	(0.496)	(0.244)	(0.271)
New hire direct team	-1.768***	-2.408***	0.366
	(0.599)	(0.358)	(0.344)
New hire spillover team	-0.263	-0.660	0.733
	(0.437)	(0.627)	(0.510)
R2	0.535	0.508	0.529
Observations	216	216	228
Skill	Low	Middle	High
Rounds	7-12	7-12	7-12
P-value (Low=Middle=High)			
AI Direct		0.09	
AI Spillover		<0.01	
New Hire Direct		<0.01	
New Hire Spillover		0.14	

Notes: This table examines heterogeneity in the effects of automation and new hires by player skills. We estimate Equation 3 on various sub-samples based on player skills. Low refers to the bottom 1/3, middle to the middle 1/3, and high to the top 1/3 of teams given each skill distribution. Columns 1–3 examine heterogeneity by the average skill distribution of the remaining players. The bottom of the table displays p-values from a joint test of treatment effect equality for low, middle, and high-skilled teams (for both automation and new hires.) All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

B.12 Validation of round difficulty measure

In section B.3 of the main text, we describe our measure of round difficulty. In this section, we validate the measure in two ways. First, we check the weights placed on each set of ingredients in the round difficulty prediction. Table B.16 displays the output

of the prediction model. The results indicate that teams are less likely to complete recipes that contain a larger number of ingredients that are farthest from the team. The model places the largest negative weights on the two recipe types that each require two ingredients that are farthest from the team. Meanwhile, the largest positive weight is placed on the recipe type that requires two of the closest ingredients to a team. This lines up with common sense priors that the closer the ingredient is to a team, the easier it is to perform the task of picking up that ingredient and bringing it to the table where ingredients are collected.

Second, we can test how well the round difficulty measure predicts performance in a given round. Table B.17 displays the results of regressions of total ingredients returned on our index for round difficulty with different fixed effects and controls in each column. Column 1 contains no controls, column 2 controls for round fixed effects, column 3 controls for round fixed effects and treatment identifiers, and column 4 controls for round fixed effects and firm fixed effects. The results indicate that teams who receive a more difficult set of recipes return fewer ingredients. This is even true when we include fixed effects for each organization: column 4 illustrates that a team returns 0.520 fewer ingredients in a round where they receive recipes that are one standard deviation higher than average, compared to that same team in a round with average round difficulty.

B.13 Co-worker preferences

In Figure B.10 we display a histogram of participant responses to the prompt: I prefer to perform this task with a human over AI. A score of 7 corresponds to strongly agree, while a score of 0 responds to strongly disagree. Human participants prefer working with humans than with AI; 84 percent reported they prefer or strongly prefer human coworkers

Table B.16: **Weights from prediction algorithm**

	(1) Total ingredients
Close-Middle-Far	0.00573
	(0.113)
Close-Middle-Middle	0.0881
	(0.104)
Close-Far-Far	-0.256**
	(0.105)
Close-Close-Middle	0.154
	(0.112)
Close-Close-Far	0.0296
	(0.106)
Middle-Middle-Far	0.0593
	(0.113)
Middle-Far-Far	-0.242***
	(0.0877)
R2	0.853
Observations	660
Round fixed effects	Yes
Firm fixed effects	Yes

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.010

Notes: This table displays the output of the prediction model. The results come from a regression of total ingredients on binary indicators for the recipe types, total recipes completed, firm fixed effects and round fixed effects. The model is limited to phase 2 of the experiment (prior to automation) and includes robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

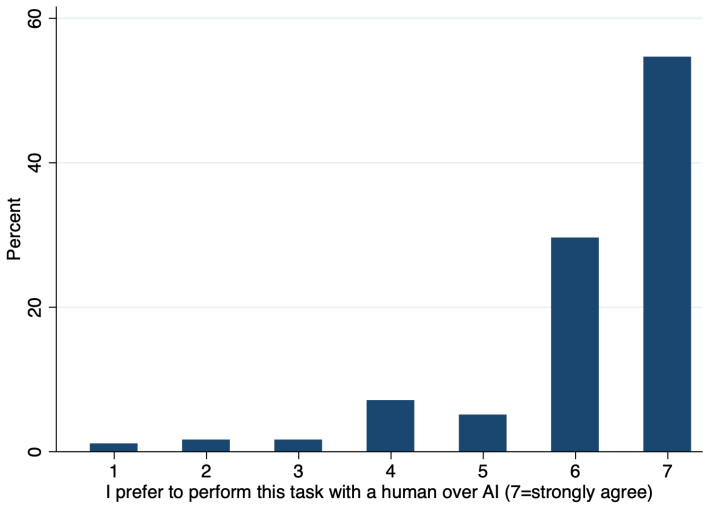
Table B.17: **Correlation between round difficulty and team performance**

	Total ingredients			
	(1)	(2)	(3)	(4)
Round difficulty	-0.600***	-0.548***	-0.547***	-0.520***
	(0.0685)	(0.0601)	(0.0604)	(0.0611)
R2	0.048	0.263	0.266	0.449
Observations	1320	1320	1320	1320
Round fixed effects	No	Yes	Yes	Yes
Treatment indicators	No	No	Yes	No
Firm fixed effects	No	No	No	Yes

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.010

Notes: This table examines the correlation between round difficulty and total ingredients returned. It displays the results of regressions of ingredients returned on our round difficulty index with different fixed effects and controls in each column. Column 1 contains no controls, column 2 controls for round fixed effects, column 3 controls for round fixed effects and treatment identifiers, and column 4 controls for round fixed effects and firm fixed effects. All regressions include robust standard errors clustered at the firm level. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Figure B.10: **Co-worker preferences**



Notes: This figure displays a histogram of participant responses to the prompt: I prefer to perform this task with a human over AI. A score of 7 corresponds to strongly agree, while a score of 0 responds to strongly disagree.