Cognitive Foundations of Collaboration

Yang Xiang[1,*], Samuel J. Gershman[1,2,3], Natalia Vélez[4]

[1]Department of Psychology, Harvard University
[2]Center for Brain Science, Harvard University
[3]Kempner Institute for Natural and Artificial Intelligence, Harvard University
[4]Department of Psychology, Princeton University
*Correspondence: yyx@g.harvard.edu

Short abstract. Collaboration enables humans to achieve goals that are beyond the reach of even the most capable individual. How does the mind solve the problem of collaboration? Here we propose the Collaborative Utility Calculus, a cognitive framework that explains how individuals make joint inferences and plan joint actions by reasoning about each others' beliefs, desires, competence, and effort. The framework captures how individuals with varying abilities redistribute effort in collaborative tasks by reasoning recursively about what their partners can do and are willing to contribute. By specifying these mechanisms, it helps elucidate how individual decisions give rise to collaborative achievements.

Long abstract. Collaboration enables humans to achieve goals that are beyond the reach of a single person by integrating the efforts of people who differ in what they know, what they want, and what they can do. This integration is at the heart of both the greatest benefits of collaboration and its greatest challenges—collaborators can disagree, they can work on the wrong things, or they can misunderstand one another. What cognitive capacities enable humans to navigate these challenges? We propose the *Collaborative Utility Calculus* as a cognitive framework for understanding human collaboration. According to this framework, humans have an intuitive theory of competence and effort, which they knit together with their intuitive theory of belief-desire reasoning to make joint inferences and plan joint actions. We present evidence for this framework from emerging work in computational cognitive science and cognitive development. These studies provide evidence that reasoning about other minds enables humans to navigate basic problems of collaboration, including deciding how much effort to invest in a collaborative task, how to incentivize collaborators to put in their fair share, whom to recruit to a team, how to invest in their training, and how to assign responsibility and rewards for the outcome of a collaboration. Finally, we discuss implications of the framework and open questions for future research.

Keywords: Collaboration, Social cognition, Intuitive theories, Theory of Mind, Joint action, Competence inference, Effort allocation, Computational modeling

1. Introduction

Two friends are lifting a couch up a flight of stairs. It's a small moment—two people aligned on a fairly simple goal, no need for elaborate tools or complicated planning. And yet in this miniature, we see all the basic components of collaboration at play. First, the two partners are *responsive* to each other's actions: if one person pauses or adjusts their grip, the other instinctively slows down or shifts position. Second, they have a *joint commitment* to completing the task. If one person stops to take a break, the collaboration does not dissolve; together, they pause, wait, and start again. Finally, they *support* each other's actions. If one partner is carrying more weight, the other might take shorter steps or lift a bit higher to rebalance the load.

Philosophical treatments of collaboration emphasize these features—mutual responsiveness, joint commitment, and mutual support—as the core criteria for shared cooperative activity (Bratman, 1992; Gilbert, 2009; Grosz & Kraus, 1996). These features allow teams to form shared beliefs (Gilbert, 1987; Tuomela, 1992) and act with shared agency. This capacity for joint action underlies many of humanity's greatest achievements—including building cities, curing diseases, and landing on the moon—and manifests across an enormous range of contexts and scales. Humans form an incredible variety of collaborative teams, from small project groups to sprawling scientific consortia, from informal family dinners to tightly organized kitchen brigades. And the tools we use to collaborate are just as varied, including email, video conferencing, physical whiteboards, and shared digital documents. Yet no matter how varied the form or how advanced the tools, human collaboration is fundamentally constrained by the limits and architecture of the human mind.

At the group level, a wide range of research has characterized how collaboration unfolds in groups and teams: how people coordinate joint actions, form shared task representations, anticipate one another's behavior, distribute roles, and collaborate in real time (Butterfill, 2012; Sebanz et al., 2006; Knoblich et al., 2011; Vesper et al., 2010; Pesquita et al., 2018; Sebanz & Knoblich, 2021; Candidi et al., 2015; Sebanz et al., 2003). These approaches have been enormously successful at describing the structure and dynamics of collaborative systems. What remains less well understood, however, is how collaborative activity is supported by cognitive representations within individual minds. How do individuals evaluate the costs and benefits of different coordination plans? How do they infer how challenging an action will be for a partner, or decide how much effort to contribute themselves? How do these local decisions that happen within individual minds accumulate and manifest at the group level, in the form of shared plans and coordinated group behavior?

In response to this challenge, several theoretical traditions have proposed key ingredients that may support collaboration at the individual level. Cognitive and developmental research has mapped out how people represent others' beliefs and desires through an intuitive *Theory of Mind* (Premack & Woodruff, 1978; Baker et al., 2009, 2017). Economic theories emphasize how cooperative behaviors are shaped by external incentives and tradeoffs (von Neumann & Morgenstern, 1944; Nash, 1950). Philosophical accounts examine the structure of shared intentions and the normative commitments they entail (Pollack, 1990; Searle, 1990; Grosz &

Kraus, 1996; Bratman, 1992). However, these accounts often stop short of explaining how the mind solves the problem of collaboration—for example, how people determine who should do what, how hard it will be for them, and whether someone is a worthwhile partner in the first place. What's missing is a framework that connects these accounts, linking theory of mind, decision-theoretic reasoning, and shared intention into a unified account of the cognitive mechanisms that support collaboration.

In this paper, we introduce the Collaborative Utility Calculus, a cognitive framework for understanding human collaboration that operates at the level of individual cognition. We first review relevant theoretical precursors, focusing on classical game theory and theory of mind. We then describe the new framework, explain how it advances existing accounts, and review recent empirical evidence for the framework. We conclude by discussing the implications of the framework for classic questions that cut across areas of cognitive science and computer science, and open questions for future research.

2. Theoretical precursors

To situate the Collaborative Utility Calculus within existing computational frameworks, we review two prominent traditions: classical game theory and classical Theory of Mind. These frameworks formalize the inferences and reasoning that support cooperation and collaboration. Understanding the contribution of these frameworks—and their limitations—sets the stage for our current proposal.

2.1 Classical Game Theory

Classical game theory is one of the most influential frameworks for modeling strategic and collective behavior (von Neumann & Morgenstern, 1944; Nash, 1950; Camerer, 2003; Smith, 1982). Game theory models how agents act optimally when the payoffs of their actions depend on others' actions. In doing so, it explains what conditions motivate people to cooperate—that is, to act together for a collective benefit rather than each selfishly pursuing their own interests. Cooperation is an umbrella term that includes not just collaboration, but also activities that do not satisfy the criteria for shared cooperative activity described above, such as donating to charity or following traffic laws. Collaboration, by contrast, is a special kind of cooperative activity in which agents work together to achieve a shared goal (Grosz & Kraus, 1996). Understanding the basic conditions that give rise to cooperation can therefore help us understand the preconditions for human collaboration.

Game-theoretic models often assume that individuals act solely to maximize their own payoffs. In a standard coordination game, there are often multiple Nash equilibria—situations where no player can obtain a higher payoff by unilaterally changing their strategy. Contrary to this assumption, however, people tend to prefer some equilibria over others, especially those that maximize the combined payoffs for everyone (Bacharach, 2006; Colman & Gold, 2018). In fact, people often seek to maximize this joint utility even when doing so reduces their own payoff (Fehr & Gächter, 2000; Fehr & Fischbacher, 2004; Bowles & Gintis, 2004) or increases their own costs (Török et al., 2019, 2021). These preferences and behaviors point to an underlying

representation of joint utility that transcends individual payoffs. One way to account for these behaviors is to modify the rewards that players receive by adding social preferences or other-regarding concerns (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000; Gallucci & Perugini, 2000; Charness & Rabin, 2002). But this addition raises a deeper question: where do these preferences come from? What internal representations support them? Although changing the payoff structure can allow cooperative strategies to emerge as equilibria, this does not by itself constitute a psychologically realistic explanation (Colman, 2024). Without a cognitive theory of how individuals construct and update these utilities, simply adding social preferences to game-theoretic models cannot provide a full explanation of cooperation or collaboration.

Theories of team reasoning take a different approach, by altering the unit of agency from the individual ("What do I want?") to the team ("What do we want as a team?") (Sugden, 1993; Gilbert, 1990). This approach can explain why people converge on collectively beneficial equilibria, but it ignores agent-specific costs. For example, suppose two equally strong people are lifting a couch together. The couch can be lifted either by one person exerting 100% effort or by both exerting 50% effort. From a team-reasoning perspective, both solutions are equally good. However, people tend to prefer the latter (Fehr & Schmidt, 1999), which reflects sensitivity not only to the outcome of the collaboration, but also to how individual costs are distributed. A related challenge concerns the formation of the team itself. Team-reasoning models typically assume that agents already reason as a team, but do not specify *ex ante* how individuals come to adopt a collective frame, or under what conditions they shift from individual to collective reasoning.

Moreover, game-theoretic models capture equilibria in stylized scenarios where the costs and payoffs to each agent are fixed. But this simple scenario underestimates how creative humans can be when they collaborate (Hertz et al., 2025; Ostrom, 1990)—for example, bringing in a more capable partner can lower the effort required from everyone. Such flexibility relies on humans' ability to represent others' goals, abilities, and effort, and use that to divide labor, train novices, assign responsibility, and respond to breakdowns in coordination. To explain how humans actually collaborate, we need a framework that captures that behavioral flexibility.

2.2 Classical Theory of Mind

A key part of what makes humans such good collaborators is their ability to reason about other people's minds, a capacity known as *mentalizing*. This ability allows humans to infer what others want, what they believe, and what they are likely to do. For example, if you see a friend straining to lift one end of a couch, you might naturally infer that their goal is to move it. This everyday inference reflects a *Theory of Mind* (Premack & Woodruff, 1978)—an intuitive understanding of how people's observable behaviors are shaped by their unobservable beliefs and desires. The ability to engage in belief-desire reasoning is core to collaboration; to coordinate with a partner, we need to understand what they are trying to achieve and what they believe about the situation.

The computations behind Theory of Mind have been formalized in Bayesian models of commonsense psychological reasoning (Baker et al., 2009, 2017), which frame action

understanding as a process of probabilistic inference over unobservable mental states. In these models, people infer what others believe and desire by observing their behavior, under the assumption that agents take efficient actions to achieve their goals. Thus, deviations from the most efficient action—such as taking a detour to check out an option hidden from view (Baker et al., 2017)—reveal the agent's goals and preferences. Extending this framework to multi-agent settings has proven fruitful for explaining group behaviors, such as how agents infer collaborators' intentions and what they should do without centralized planning (Wu et al., 2021), and how agents use other people's behavioral trajectories to learn about the environment when reward information is private (Hawkins et al., 2023). These models illustrate how probabilistic mental-state reasoning can scale to more complex social environments. However, these models assume a universal notion of efficiency by treating action costs as fixed. In reality, what counts as an "efficient" action depends on individual differences in ability and effort.

The Naive Utility Calculus framework extends Bayesian Theory of Mind models by incorporating individual variation in costs and rewards (Jara-Ettinger et al., 2016); for example, climbing a tall box may be easy for one agent but impossible for another (Jara-Ettinger et al., 2015). However, in collaborative settings, the cost to individuals depends not only on their own capabilities, but also on what their collaborator contributes. Climbing is easier if your partner gives you a boost from below, compared to if your partner is barely trying to help. This gap motivates the need for a framework that captures how collaborators redistribute effort to achieve shared goals.

3. The Collaborative Utility Calculus

In this section, we propose the *Collaborative Utility Calculus* as a cognitive framework for understanding human collaboration. The framework makes unique contributions to existing approaches. It explains collaboration from the perspective of basic mechanisms of human social cognition, which includes concepts such as mental representations of joint utility that are not typically captured in classical game theory. The framework also extends existing models of commonsense psychological reasoning into the domain of collaboration to explain how people reason about joint utility and plan shared actions. It makes explicit theoretical commitments about how *competence* and *effort* jointly determine an agent's contribution to a collaborative task. It also formalizes the idea that people anticipate how work will be distributed in joint settings, and that how much effort one exerts depends in part on how much their partner is expected to contribute.

Figure 1 shows an illustration of the framework. The framework includes a number of agent-specific properties: (a) the reward value of an action or outcome (*desire*), (b) what the agent knows about themselves and the world (*beliefs*), (c) the agent's ability to carry out the task (*competence*), and (d) how hard the agent tries at the task, i.e., the proportion of their competence applied to the task (*effort*). These properties produce an action, which is compared with the *task difficulty* to determine whether a collaboration succeeds or fails (*outcome*). The outcome then feeds back into the agent's beliefs: it becomes part of what the agent knows and is used to adjust their future actions. See Box 1 for a formal description of the framework.

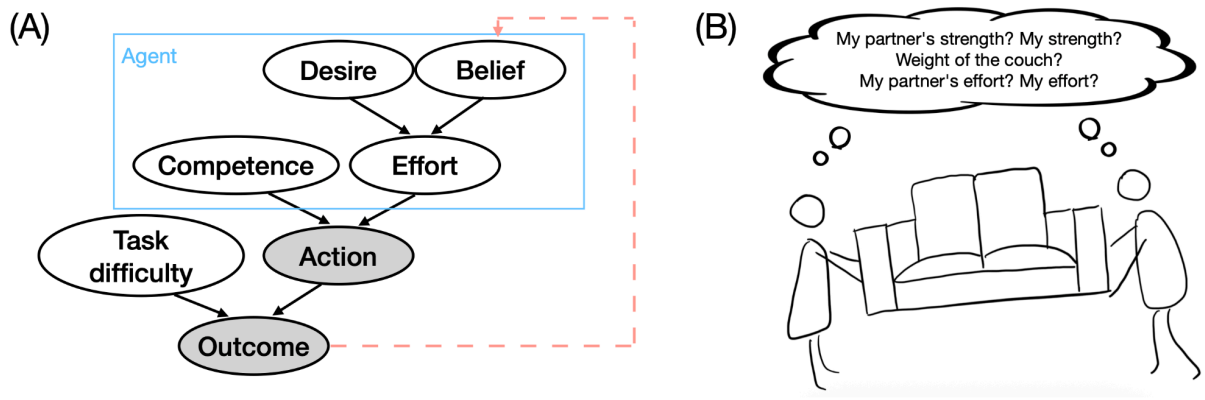

Figure 1. Overview of the Collaborative Utility Calculus. (A) Schematic of the Collaborative Utility Calculus, (B) illustrated in an example of two friends lifting a couch together. Each person considers their own mental states and competence, as well as their partner's mental states and competence, and chooses actions that maximize the joint utility of the team.

Box 1: Formal modeling of the Collaborative Utility Calculus

We consider agents working together on a task with difficulty D. Each agent (denoted by a) has competence S_{a} and exerts effort defined as a proportion of their competence applied to the task. This produces exertion $F_{\mathrm{a}} = S_{\mathrm{a}} \times E_{\mathrm{a}}$. Here, we use collaborative box-lifting as an illustrative example. In the context of box-lifting, competence refers to an agent's strength, and exertion refers to the force they exert.

For simplicity, we assume that agents can lift the box when their combined force exceeds the box weight, i.e., $\sum_{a} F_{\mathrm{a}} \leq W$. To map these terms onto the framework, strength is their competence, force is their action, and box weight is the task difficulty.

Each agent gets reward R_{a} if and only if the team succeeds. The agents' strength is unknown, so the probability of success is given by marginalizing over each agent's strength:

$$P(\text{success}|\mathbf{E}, W) = \int_{\mathbf{S}} P(\mathbf{S}) \cdot \mathbb{I}[\sum_{a} F_a \geq W] d\mathbf{S}, \quad (1)$$

where $\mathbf{E} = [E_1, \ldots, E_n]$ denotes the vector of efforts for all n agents and $\mathbf{S} = [S_1, \ldots, S_n]$ denotes the vector of strengths. Here, we use $P(\mathbf{S})$ generically to denote beliefs about agents' strength distribution, which could either be a prior belief or a posterior belief conditional on the evidence.

The utility of exerting effort is defined as the net reward (i.e., expected reward minus cost):

$$U(\mathbf{E}) = P(\text{success}|\mathbf{E}, W) \sum_{a} R_a - C(\mathbf{E}) \quad (2)$$

where C(**E**) is an umbrella term for all the costs agents incur in the process, which can include each agent's effort cost and social costs such as a penalization for unequal effort allocations.

We model agents as utility maximizers, thus the optimal effort **E** is given by:

$$\mathbf{E}^* = \underset{\mathbf{E}}{\operatorname{argmax}}\, U(\mathbf{E}) \quad (3)$$

In practice, we solve the joint optimization problem using fixed point iteration, by maximizing each agent's utility with respect to their own effort while keeping the efforts of all other agents fixed, iterating over agents until the solution converges.

Note that the framework applies more broadly to agent properties that are conceptually similar to competence and effort, beyond the literal sense. For example, a person's available resources or wealth can be thought of as analogous to competence—that is, the maximum they can contribute, whereas their generosity or fairness maps onto effort, as it reflects the proportion of their capacity they choose to allocate to their partner (Hackel et al., 2015; Raihani & Barclay, 2016). The framework can also capture situations where factors other than effort might reduce how much competence an agent brings to a task, such as temporary limitations (e.g., fatigue or illness), environmental constraints (e.g., bad weather or a noisy room), or even strategic self-handicapping (Xiang et al., 2026).

All these nodes interact with one another in intricate ways. First, competence and effort work together to determine the outcome of a collaboration. Even the most competent agents cannot succeed without investing the required effort, nor can the most hardworking agents if they lack the required competence. Second, agents choose effort allocations that maximize their joint utility—the expected reward for the team minus the costs of exerting effort. This ensures that agents prioritize the team's shared goal while also being sensitive to the individual costs they each incur. Third, the expected reward calculation hinges on joint inferences about one's own and their collaborators' competence and mental states. For example, if you are lifting a heavy couch with a friend, it is important to understand how strong your friend is—i.e., their competence—and how much effort they will put in, so that you can calibrate how much effort to apply to lift your end of the couch. The framework predicts that agents recursively reason about how much effort they and their collaborators will allocate to a task. With these elements, the framework can capture a wide range of phenomena in human collaboration, as we review in the next section.

4. Evidence for the Collaborative Utility Calculus

In this section, we review recent empirical evidence for the Collaborative Utility Calculus as a cognitive framework for understanding human collaboration. These studies collectively show how reasoning about other minds enables humans to navigate basic problems in collaboration, including deciding how much effort to invest in a collaborative task, how to incentivize collaborators to put in their fair share, whom to recruit to a team, how to invest in their training, and how to assign responsibility and rewards for the outcome of a collaboration. In many of the studies below, the Collaborative Utility Calculus was compared against simple alternative

models to demonstrate the need to reason recursively about collaborators' mental states and competence. In addition, we highlight the potential for this framework to shed light on a wide range of phenomena, such as how people reason about external and internal constraints, how collaborators achieve efficient division of labor, and how collaborations evolve over time.

4.1 Inferring competence and effort

In an episode of *30 Rock*, Kenneth Parcell created a game show where models held up identical-looking cases and contestants guessed which one contained $1 million worth of gold. The show was quickly shut down after it became clear that the design had a fatal flaw: since gold was heavy, contestants could easily identify the correct case by noticing which model was visibly struggling to hold it up. Underlying this funny scene is a basic principle of social reasoning: people can infer others' effort and competence—in this case, because the struggling model is not obviously weaker than the other models, one can infer that her case is unusually heavy. This ability is also core to collaboration: to work together effectively, we need to understand what our collaborators can do and how much effort they are willing to contribute.

Even children are capable of inferring other people's competence based on the difficulty of the tasks they are willing to take on. In one study, 5–6 year olds watched two puppets—Grover and Cookie Monster—take a cracker that was placed on a short box rather than a cookie that was placed on a tall box. Children were told that Cookie Monster likes cookies more than crackers, while Grover likes both equally. From this scene, children inferred that Cookie Monster was *unable* to climb the box—if he could, he would have taken the cookie (Jara-Ettinger et al., 2015). Even in cases where task difficulty is unknown, adults can jointly infer task difficulty and other people's competence by observing which agents succeed at more tasks and which tasks have higher success rates (Jara-Ettinger & Gweon, 2017). The Naive Utility Calculus explains how people make these inferences by reasoning about subjective rewards and costs—that is, how much an agent values each goal and how hard it is for that agent to reach it.

The Collaborative Utility Calculus builds on this foundation in two ways. First, it accounts for *effort* as a component of subjective cost that is distinct from ability—recognizing that people may fail to reach a difficult goal not because they lack the ability, but because they are unwilling to expend effort or because their collaborator did not put in their fair share. In one study, we showed participants two game show contestants who attempted to lift boxes of different weights for different reward amounts (Xiang et al., 2023a). People inferred which contestant was stronger not just from their successes or failures, but also from how these outcomes related to the *incentives* offered to contestants. For example, when both contestants successfully lifted a heavy box, participants inferred that the contestant who succeeded with a smaller reward was stronger—intuitively, stronger contestants require smaller rewards to offset the costs of lifting. Consistent with this reasoning, participants also inferred that stronger contestants exerted less effort to lift the same weight (see Figure 2A-2B).

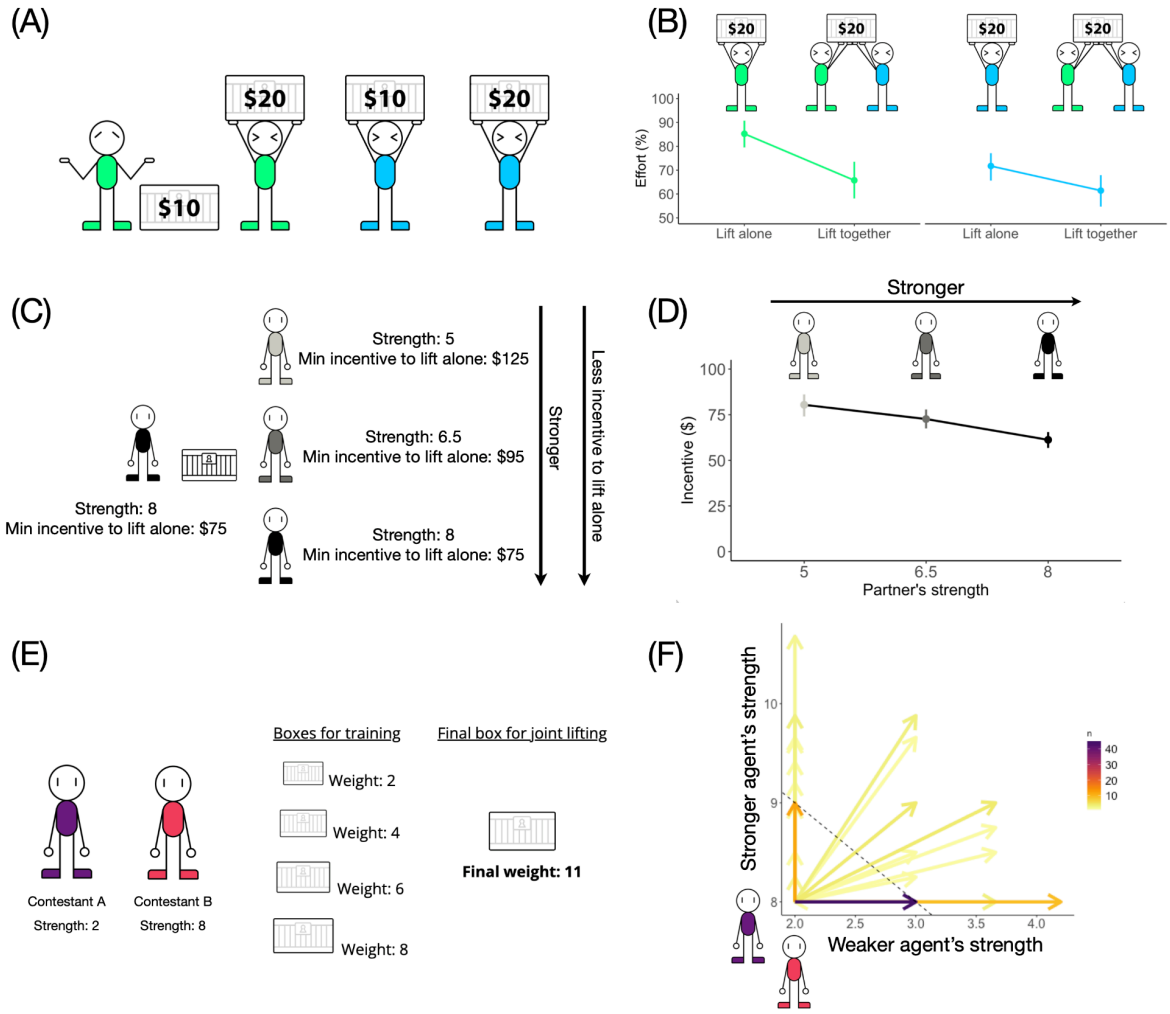

Figure 2. Empirical evidence for the Collaborative Utility Calculus. (A-B) Effort allocation (Xiang et al., 2023a): People can infer agent-specific costs—namely, competence and effort—based on the rewards that agents are willing to accept to do a difficult task. When agents lifted a heavy box alone, participants inferred that agents who lifted the box for a small, $10 reward (blue agent) were stronger and exerted less effort than agents who only lifted the box for a large, $20 reward. When the agents lifted the box together, participants inferred that each agent would exert less effort than when they lifted on their own. **(C-D) Incentive allocation** (Xiang et al., 2023a): The Collaborative Utility Calculus enables inferences about how one person's effort may change based on their partner's contributions. Participants offered smaller incentives when the agent on the left was paired with a stronger partner, reflecting the expectation that stronger teams expend less effort to complete the task. **(E-F) Optimizing competence** (Xiang et al., 2024): The Collaborative Utility Calculus also supports more sophisticated planning based on how collaborators' competence may change with training. Participants trained agents to be just strong enough to lift a heavy box together through their combined strength, balancing the costs and benefits of training.

Second, the Collaborative Utility Calculus goes beyond inferring utilities from *individual* actions to disentangling individual utilities from *joint* actions, where multiple people's contributions are combined into a single outcome. In the same study as above (Xiang et al., 2023a), participants then saw pairs of contestants attempt to lift a box together. They inferred that each contestant exerted less effort when sharing the weight with a partner, compared to lifting it alone—suggesting that participants understood that collaborations distribute effort between

partners (see Figure 2A-2B). Participants' judgments were captured by the Collaborative Utility Calculus, which models how each contestant adjusts their effort by reasoning recursively about how much their partner will contribute, but not by simpler heuristic models that lacked recursive reasoning.

Thus, the Collaborative Utility Calculus captures human inferences about competence and effort in collaborative tasks. A key question for future work is whether this same reasoning also governs how people calibrate their *own* effort in collaborations. This possibility aligns with recent developmental work suggesting that young children use social information to estimate the difficulty of a task and calibrate their effort accordingly. For example, infants pull harder with each subsequent attempt to retrieve a toy that is difficult for an adult to move, but do not increase their effort if the task was impossible for the adult (Lucca et al., 2020). Infants also try harder after observing an adult try repeatedly at a different task, rather than when the adult succeeds effortlessly (Leonard et al., 2017). Critically, infants do not indiscriminately copy adults' effort; they selectively increase their own efforts when they see that adults' struggles lead to success, rather than failure (Leonard et al., 2020). These social cues provide infants with information about how effortful a task will be *for them*, and whether their effort is likely to pay off. Applying the Collaborative Utility Calculus to this domain could shed light on the psychological mechanisms that enable people to translate social information about other people's capabilities and outcomes into principled decisions about when and how much to contribute.

4.2 Using competence and effort for collaborative decision-making

A key implication of the Collaborative Utility Calculus is that reasoning about other people's competence and effort guides practical decisions about how to build a team, by revealing what team members can do and what costs they are expected to incur. This information can then be used to incentivize collaborators to put in their fair share. In one study, we asked participants to choose a monetary incentive that both team members would receive if they successfully lifted a heavy box together (Xiang et al., 2023a). Across conditions, participants saw the same agent paired with partners of different strengths. Participants offered lower incentives when this agent was paired with a stronger teammate, and higher incentives when paired with a weaker teammate (Figure 2C-2D). These results suggest that participants set incentives to compensate for the *effort* that teams would have to put in to complete the task. Intuitively, stronger teams expend less effort to produce the same physical force as weaker teams, so they require smaller incentives to make the task worthwhile. Similarly, when we increased one agent's strength, participants reduced the team's incentive accordingly. These patterns were captured by the Collaborative Utility Calculus, but not by alternative models that use simple heuristics in place of recursive reasoning.

The cost of effort can also vary depending on individual differences in each collaborator's *willingness* to contribute. In another study, participants recruited pairs of agents out of a lineup (Xiang et al., 2023a). Each agent in the lineup differed not only in their strength, but also in their *effort cost*—that is, the lowest incentive that they would accept to put in an all-out effort. Intuitively, agents with lower effort costs are "hard-working", or more willing to exert effort. After picking a team, participants then offered each team member an incentive to lift boxes of varying

weights. This design enables us to test how participants build teams by weighing each agent's strength against their willingness to expend effort. When the box was light, participants recruited weaker but more hardworking agents who required a smaller incentive to expend the necessary effort. In contrast, when the box was too heavy for the weaker agents, participants recruited stronger but less hardworking agents and were willing to offer larger incentives.

Willingness to exert effort matters not only for the success of a single collaboration, but also for identifying people who will be good collaborators in the future. In another study, participants selected how much each agent in a pair should be rewarded for successfully lifting a box together (Xiang et al., 2025a). Participants were told that one agent was *instructed* to exert a fixed amount of effort, while the other chose how much effort to invest on their own. Participants apportioned rewards based not only on how much effort each agent had invested, but also on their reasons for doing so. Overall, participants gave agents rewards that were in line with their expended effort, but this effect was moderated by whether they were instructed or self-directed. They gave the self-directed agent rewards that were more proportional to actual effort expended, compared to the instructed agent. These results suggest that people evaluate collaborators based not only on effort, but also on what that effort reveals about their willingness to contribute in future collaborations.

Finally, reasoning about competence and effort is crucial for evaluating the outcome of a collaboration—that is, to decide who deserves a larger share of blame for a failed collaboration or credit for a successful one. Past work has shown that people are sensitive to both effort (Bigman & Tamir, 2016; Jara-Ettinger et al., 2014) and competence (Gerstenberg et al., 2011) in responsibility judgments. However, these factors have largely been studied in isolation, which leaves open the question of how people integrate competence and effort to assign responsibility. In a set of studies, participants observed pairs of agents lifting a box together and evaluated how responsible each agent was for the outcome. We systematically varied box weight, agent strength, and effort to test how each factor contributed to participants' responsibility judgments. Overall, participants assigned responsibility based on effort, rather than competence or the amount of force exerted on the box. Moreover, participants' judgments were shaped by how much effort each agent *actually* contributed and by a counterfactual judgment of how much the agent *could have* changed the outcome by exerting more or less effort, given their competence. Intuitively, the agents who received the most credit for successes exerted a large amount of effort *and* were pivotal to their team's success (Xiang et al., 2023b).

These studies collectively show how reasoning about other minds enables humans to navigate basic problems in collaboration, including deciding how much effort to invest in a collaborative task, how to incentivize collaborators to put in their fair share, whom to recruit to a team, and how to assign responsibility and rewards for the outcome of a collaboration. A natural next step is to consider how the Collaborative Utility Calculus may inform decision-making as collaborations unfold. One intriguing possibility for future work is that our framework may provide a psychological mechanism to explain how human collaborations achieve efficient divisions of labor. As collaborators reason recursively about each other's competence and effort, they may converge on efficient task allocations through mutual adaptation. This proposal aligns with

recent empirical work on “co-efficiency”, showing that collaborators often act to minimize joint costs (Török et al., 2019, 2021; Sebanz et al., 2006; Knoblich et al., 2011; Brennan et al., 2008).

4.3 Optimizing competence

The sections above have established that, in order to efficiently divide labor with others, we need to represent our collaborators’ competence. However, competence is not static; people get better at particular tasks the more they perform them. Indeed, this opportunity to improve has long been touted as a key benefit of division of labor: splitting up tasks affords individuals more chances to practice and refine their area of specialization (Smith, 1937). However, it also raises an important question in collaborative decision making: Is it better to recruit a superstar who is the most capable right now, or a rising star whose competence can grow with training and investment? And how do decisions about whom to train affect whom people choose to recruit in the first place?

According to the Collaborative Utility Calculus, people make these decisions by weighing the costs of recruiting and training a collaborator against the benefits that they will bring to the collaboration. In one study, participants were shown a heavy box that required two agents working together to lift successfully. They then had to recruit agents of varying strengths from a lineup and train them to reach the combined strength needed for this collaborative goal (Xiang et al., 2024). We found that participants did not simply recruit the strongest agents for the job, nor invest in those who would improve the most. Instead, they strategically selected and trained agents to be *just strong enough* to lift the target weight, balancing the costs of training against the benefits (see Figure 2E-2F). These results generalized to a separate experiment where participants trained students for a math Olympiad, which suggests that these training decisions generalize beyond increasing physical strength. In another study, participants trained two military defense teams to counter two types of attacks (land and air), where the long-term goals and the teams’ relative competences varied (Tian et al., 2026). Overall, participants trained collaborators by taking the long view—considering how collaborators’ competence would develop, and whether they could rise to meet task demands—rather than making myopic decisions based on simpler heuristics such as current competence, learning potential, fairness, or versatility. Together, these findings illustrate how the Collaborative Utility Calculus captures not just static inferences about others’ abilities, but also how people cultivate individual competences to support collaboration over time and ultimately benefit from each other’s evolving competences.

5. Relation to group-level processes

The Collaborative Utility Calculus is an individual-level account of collaboration: it specifies how individuals represent costs, rewards, competence, and effort when collaborating with others. Collaboration has also been studied at broader levels of analysis, including dyads (Sebanz et al., 2006; Knoblich et al., 2011; Brennan et al., 2008; McCarthy et al., 2021; Bahrami et al., 2010), small groups (Larson, 2013; Almaatouq et al., 2021; Bernstein et al., 2018; Woolley et al., 2010), and even joint actions spread over massive networks such as care teams (Edmondson, 2012),

ship crews (Hutchins, 1995), intelligence networks (Hackman, 2011), and self-governing online communities (O'mahony & Ferraro, 2007; Vélez et al., 2024). Scaling up individual-level theories of collaboration to groups introduces new challenges, including understanding how collaborators align on shared representations and what institutional and social tools they have available to maintain motivation among the group. In this section, we situate the Collaborative Utility Calculus within this problem space, clarifying how characterizing individual-level cognitive mechanisms may provide new ways to understand and intervene on group-level behaviors.

5.1 Understanding social motivations to collaborate

Why do people commit to collaborative goals in the first place? Many collaborative activities cannot be explained solely in terms of immediate instrumental benefits to all participants. Individuals often collaborate when acting alone is more effective to achieve the goal (Curioni et al., 2022), and even young children expect agents to collaborate in situations where the same outcome could be achieved by themselves with less costs (Vizmathy et al., 2024). These patterns raise a foundational puzzle: What motivates people to work together, even when collaboration does not bring additional benefits to the individual or entails additional costs?

One influential answer comes from social economics, which argues that individuals' preferences extend beyond their own material payoffs: people may also derive utility from others' welfare (Charness & Rabin, 2002), from fair distributions of payoffs (Fehr & Schmidt, 1999), or from the reputational benefits of cooperation (Nowak & Sigmund, 2005). By incorporating these social concerns directly into the utility function (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000; Gallucci & Perugini, 2000; Charness & Rabin, 2002), seemingly altruistic collaboration can remain instrumentally rational: individuals collaborate because doing so maximizes a broader conception of utility that includes social goals.

The Collaborative Utility Calculus takes a different approach by endogenizing certain forms of social information. For example, rather than adding a term for other-regarding preference in individual utility functions, the framework shows how this preference can arise from maximizing one's immediate benefits in an interdependent setting, where success requires everyone to pull their weight. Beyond describing social concerns by adding new preference parameters, this approach provides a mechanistic explanation for how certain cooperative behaviors emerge from individual social reasoning.

5.2 Coordinating joint actions

How do individuals coordinate their actions in real time to achieve shared goals? Even in simple dyadic tasks—lifting a couch, dancing a waltz, passing a basketball—successful collaboration requires partners to anticipate one another, align their behaviors, and adjust dynamically as circumstances change. A large body of research on joint action has shown that collaborators form shared representations of the task and of each other (Vesper et al., 2010; Knoblich et al., 2011), and use these representations to infer joint commitments (Michael et al., 2016) and predict upcoming actions (Sebanz & Knoblich, 2021; Sebanz et al., 2006). This line of work identifies a core principle of shared action plans: they tend to minimize collective costs, reflecting sensitivity to joint efficiency (Török et al., 2019, 2021; Mascaro & Csibra, 2022).

While this literature provides a rich characterization of the structure and dynamics of coordinated behavior, it does not fully explain how individual agents arrive at a shared action plan, particularly when costs differ across collaborators. What internal trade-offs are evaluated when deciding how much effort to exert? How do differences in competence or willingness to bear costs shape the division of labor? The Collaborative Utility Calculus complements joint action research by specifying how individuals represent costs, rewards, competence, and effort within a unified utility-based framework. In doing so, it highlights that the best solution for a team depends on how costly a given action is for each collaborator. These asymmetries can be built into real-time coordination tasks to better understand how people divide work when they differ in ability or in how much effort they are willing to contribute.

5.3 Embedding collaboration within cultural and institutional structures

In recent years, humans have been able to collaborate at a scale that is unprecedented in our evolutionary history. We can now pursue projects that span decades and involve thousands of collaborators spread all over the world. For example, the ATLAS collaboration—a large particle physics project at CERN involved in the discovery of the Higgs boson—involves more than 5000 researchers worldwide supported by massive technological and institutional infrastructure (ATLAS collaboration, 2022). These large-scale collaborations have been uniquely enabled by both technological innovations that facilitate communication and by institutional innovations to fairly distribute tasks, decisions, and credit among the group. Yet collaborating at this grand scale is not trivial. For every success story enabled by these technological and institutional structures, there are also untold numbers of teams floundering under unclear direction, clashing priorities, and ineffective communication. What makes large-scale collaborations possible, and why do they sometimes fail?

One influential approach to this question treats collaboration as a *distributed cognitive system* embedded in cultural and institutional structures. Theories in distributed cognition and team cognition argue that collaborations include not just the minds of individual collaborators, but also structural supports external to individual minds—such as institutional roles, routines, and material artifacts (Hutchins, 1991, 1995; Fiore et al., 2010; Wegner, 1987; Stahl, 2006; Cooke et al., 2013; Cooke, 2015). Hutchins' (1995) ethnography of Navy ship navigation provides a canonical example: the computational problem of navigation is solved by combining the efforts of individual crew members with a suite of navigational tools and communication protocols. No single crew member represents "the whole plan" in their mind. In this sense, teams can function as cognitive systems whose operations emerge from structured interaction rather than from centralized planning in any one individual. More broadly, distributing cognition across individuals and across generations also allows for rapid and transformative cultural evolution. The body of cultural knowledge is too vast for any single individual to possess (Gabora, 2013), and technology can be too complex for individuals to operate on their own (Miton & Jackson, 2025). By dividing the cognitive labor, each individual only needs to learn a subset of knowledge or skills and can act as a node in a vast distributed system that enables collaborations at scale (Vélez et al., 2023).

The Collaborative Utility Calculus complements these system-level accounts by specifying the individual-level reasoning that operates within such structures. It helps explain why even collaborators embedded in the same institutional structure may behave differently due to differences in their abilities or costs. For example, two sailors occupying the same monitoring role may adopt different reporting thresholds depending on differences in expertise, fatigue, or incentives from their supervisor. Distributed cognition frameworks describe how roles and artifacts structure coordination, but they typically do not model how asymmetric costs translate into individual effort allocation or decision thresholds. The Collaborative Utility Calculus helps fill this gap by modeling how individuals reason about costs within these larger structures.

6. Implications

So far, we have shown that the Collaborative Utility Calculus expands the expressive power of Bayesian mentalizing models to capture a wide range of decisions that are central to collaboration, such as allocating effort, building a team, and investing in training team members. Our modeling framework grounds collaboration in basic mechanisms of human social cognition, and is thus compatible with emerging work in computational cognitive science and cognitive development. In this section, we examine the developmental origins of these cognitive capacities, their implications for pedagogy and communication, and their role in supporting large-scale collaborations. We conclude by considering how the Collaborative Utility Calculus may provide a principled basis for comparison between human collaborations and emergent behaviors in artificial systems.

6.1 Developmental Emergence of the Collaborative Utility Calculus

The Collaborative Utility Calculus builds on three key ingredients that emerge early in development: an understanding of action efficiency, a motivation to act prosocially and jointly with others, and the ability to reason about one's own and others' competence. We trace how each of these components develops during infancy and early childhood to form the foundation for collaborative decision-making. There are some parallels of these abilities in non-human primates, though understanding whether these capacities are innate inductive biases conserved across species and cultures remains an exciting direction for future work (see Box 2).

The first key ingredient is an early-emerging understanding of action efficiency—the expectation that agents minimize the costs they incur to achieve their goals. This principle is foundational to the Collaborative Utility Calculus, which assumes that agents weigh effort costs against rewards to make decisions. To support this reasoning, infants must first learn to interpret others' actions not just as movements, but as goal-directed behaviors. By 9 months, infants interpret others' actions as goal-directed: for example, they understand that an experimenter reaching for a toy intends to grab that toy—rather than to reach for that location—and are surprised when the experimenter continues reaching towards that location after the toy is moved. This suggests that they understand the experimenter's action as directed toward a specific goal object rather than a spatial location (Woodward, 1998). This understanding is accompanied by expectations about how rational agents should act. Infants as young as 12 months expect agents to take efficient

paths to a goal. They are not surprised when an agent hops over a wall to get to a goal but are surprised when the agent makes unnecessary detours, such as hopping over a clear path (teleological stance; Gergely & Csibra, 2003).

This understanding of action efficiency depends on infants representing other minds and bodies interacting with the physical world. That is, infants interpret actions as a product of an agent's internal properties (e.g., their intent) and external constraints (e.g., the location of the goal) (Liu et al., 2025). This understanding is supported by two key insights. The first is an understanding of the physical demands of goal-directed action. Past work has found that training 3-month-old, prereaching infants to grab objects with "sticky mittens" enables them to better recognize when other people use inefficient means to reach for an object, compared to infants who do not have this early training (Sommerville et al., 2005; Skerry et al., 2013; Gerson & Woodward, 2014). The second is an early-emerging understanding of people as causal agents. Prereaching infants are surprised when someone uses inefficient means to light up a toy—which causes a visible change—but not when they use inefficient means to simply lift the toy without changing its state (Liu et al., 2019).

By the end of the first year, infants can flexibly apply this understanding of action efficiency to predict how much effort an agent will exert to achieve a goal, and conversely, to infer how much an agent values a goal based on the costs they are willing to incur (Liu et al., 2017). These inferences indicate that infants do not simply learn statistical associations between actions and outcomes; rather, they form a generative model of agents as intentional actors. This capacity is foundational to the Collaborative Utility Calculus, which posits that effort is not a generic cost, but rather emerges from the interaction between internal properties and external constraints. The early emergence of this generative model suggests that, from infancy, children are equipped to use effort as a window into others' minds.

The second ingredient is a basic set of capacities and motivations that support collaborative action. The first is an early-emerging motivation to help: toddlers will help others achieve their goals, even when they receive no immediate benefit and the person helped is a stranger (Warneken & Tomasello, 2006; but see Cortes Barragan & Dweck, 2014). The second is *shared intentionality*, a motivation to follow through on shared goals—instead of viewing others as tools for achieving their own goals, children commit to a joint activity and view themselves and their partners as bound by an obligation to complete it (Tomasello & Carpenter, 2007). When this obligation is not met—for example, when a collaborator stops participating—toddlers make attempts to re-engage them (Warneken et al., 2006). Third, children grasp the interdependent nature of collaborative tasks: they understand that one person's contribution depends on what another person does (Henderson & Woodward, 2011) or can do (Warneken et al., 2014), and that success often requires coordination between the two parties (Brownell et al., 2006). Together, these capacities create the foundation for the Collaborative Utility Calculus, enabling children to move beyond individual goal pursuit toward coordinated action with others.

The third ingredient is the ability to reason about one's own and others' competence. While infants show early sensitivity to action costs, reasoning about competence seems to emerge

later in childhood. 4- and 5-year-olds understand that a person is more competent than their partner if they complete the same task faster, or complete a more difficult task within the same time (Leonard et al., 2019). By age 5, children begin dividing physical and cognitive labor strategically, assigning easier tasks to less competent partners and harder tasks to more competent partners (Magid et al., 2018; Baer & Odic, 2022). Younger children struggle with these judgments (Baer & Odic, 2022), suggesting that this ability requires more advanced social-cognitive reasoning. The ability to assess differences in competence and divide labor accordingly represents an important step towards the Collaborative Utility Calculus, which depends on understanding both individual abilities and how to combine them effectively.

The Collaborative Utility Calculus adds a crucial fourth ingredient: an understanding of effort in relation to one's competence. While past developmental studies have largely examined the two separately, with effort often treated as a general cost term, the Collaborative Utility Calculus highlights the need for a richer account. It explains why individuals with similar competences may contribute very different amounts of work—due to differences in their willingness to exert effort—and how effort is redistributed across team members. Thus, a promising direction for future research is to investigate how children come to represent competence and effort as distinct but interacting components, and how they use these representations to guide collaborative decisions.

Box 2: Comparative aspects of collaboration

Comparing collaboration across species and cultures reveals that there is no single formula for working together. While human collaboration is often characterized by centralized organization, other species sometimes solve coordination problems through decentralized structures. For example, ants collectively solve geometric puzzles like the piano-movers puzzle by locally adjusting to one another without representing the geometry of the problem, which allows for efficient scaling of cognitive capacities and leads to more robust performance in larger groups (Dreyer et al., 2025). Honeybees similarly use decentralized decision-making to choose among nectar sources (Seeley et al., 1991). These decentralized strategies reduce communication demands and facilitate efficient scaling of cognitive capacities, especially valuable given the cognitive limitations of a single ant or honeybee. By contrast, human collaborations often involve centralized structures (e.g., selecting leaders), which can improve flexibility but also introduces bottlenecks for large groups—groups typically perform as well as the best individual leader, and lose many of the benefits of aggregating many people's actions. These comparisons help reveal the tradeoffs of different collaborative structures.

Human collaborations also differ across cultures. For example, American schoolchildren tend to assign leaders and distribute tasks, while Mexican-heritage children often coordinate fluidly without explicit leadership (Alcalá et al., 2018). Even within Western cultural contexts, there is variation in the extent to which people emphasize independence or interdependence along class lines. For example, U.S. students from working-class communities—which tend to emphasize interdependence—perform better when they work in groups, compared to working individually (Dittmann et al., 2020). These findings suggest that cultural contexts shape

collaboration styles, and success depends on whether the metrics used to assess achievement align with these cultural styles.

Together, these cross-species and cross-cultural perspectives suggest that there is no one-size-fits-all solution to collaboration, but rather a flexible set of different strategies tuned to different ecological and social demands. They also highlight a key tradeoff in human collaboration: our ability to mentalize allows for flexible, role-sensitive coordination, but often at the cost of simplicity and scalability. This raises exciting possibilities for extending the Collaborative Utility Calculus to not only model how individuals reason about collaboration, but also what we can learn from the broader repertoire of collaborative strategies observed across species and societies.

6.2 Pedagogy, communication, pragmatics

As we have seen, the Collaborative Utility Calculus captures how people incentivize collaborators to exert effort and train them to accomplish goals just beyond their current capabilities. These properties have important implications for understanding how people teach and acquire complex skills in collaborative relationships—such as between advisor and student, or artisan and apprentice—and thus have the potential to enrich existing theories of pedagogy and communication.

Bayesian models of cooperative communication formalize pedagogy as a series of recursive inferences between a teacher and learner. A key implication of this principle is that the cognitive processes underlying teaching and social learning are two sides of the same coin (Gweon, 2021): The teacher (or speaker) selects information to maximize the learner's belief in a target hypothesis, while the learner (or listener) interprets the information to infer the target hypothesis. This process can be described using mathematical theories of *optimal transport*, where the goal of communication is to efficiently transfer beliefs into the listener's mind while minimizing costs of providing information (Wang et al., 2020; Shafto et al., 2021).

This principle explains an impressive range of communicative behaviors. Consistent with this framework, teachers strategically select information that is tailored to the learner's beliefs and goals (Chen et al., 2024; Shafto et al., 2014; Vélez et al., 2023), while learners can actively draw inferences that go beyond what teachers explicitly demonstrate (Bonawitz et al., 2011) to evaluate whether teachers are reliable, knowledgeable, and engaged (Shafto et al., 2012; Sobel & Kushnir, 2013; Bass et al., 2024). These teaching and learning behaviors emerge across a variety of modalities, including learning from examples (Shafto et al., 2014), demonstrations (Ho et al., 2021), advice (Vélez & Gweon, 2019), and verbal descriptions (Sumers et al., 2023). In the domain of language, this principle is related to the *rational speech act* model (Goodman & Frank, 2016; Degen, 2023), a theoretical framework that explains many instances of flexible language use, including hyperbole, generics, and polite speech (Yoon et al., 2020; Tessler & Goodman, 2019; Kao et al., 2014).

Incorporating representations of competence and effort into Bayesian models of communication and pedagogy opens the door to addressing deeper puzzles about how we teach and communicate. Prior empirical work suggests that even young children can identify competent individuals. Preschoolers selectively learn from reliable teachers and understand that competence is context-specific (Koenig et al., 2004; Sobel & Kushnir, 2013). For example, they recognize that other children can be better informants than adults about which toys are most fun (VanderBorght & Jaswal, 2009), and that people who are experts in one domain might not be experts in another (Lutz & Keil, 2002; Keil et al., 2008). Yet, to date, little work has considered how competence changes over the course of teaching, or how competence and effort interact. Communicating with someone who is highly competent but disengaged, or highly motivated but incompetent, calls for different strategies, yet current models rarely distinguish between these cases (but see Bass et al., 2024). In addition to addressing these moment-to-moment challenges, the Collaborative Utility Calculus offers an account of how humans develop and nurture each others' competence through long apprenticeships or mentorships—for example, deciding what to teach by considering not only how to convey the most information at the moment, but also anticipating how a learner's competence might evolve with practice, and what kinds of challenges are appropriate at different stages (Xiang & Gershman, 2025b; Chen et al., 2024; Ham et al., 2025).

6.3 Large-scale collaborations

The Collaborative Utility Calculus relies on individuals reasoning about others' mental states and abilities. But in large-scale collaborations, these processes—reasoning about what each of our collaborators thinks, and what they think other collaborators think—can quickly become computationally intractable. How do these individual-level principles scale up to reasoning about many minds over an extended collaboration?

This problem is particularly salient in larger collaborations. Suppose that you are a theorist collaborating with a molecular biology lab that runs experiments. Tracking the mental state of each biologist individually may quickly become intractable. One way to simplify the process of reasoning about other minds is to abstract away from individuals and to treat whole groups as a single entity. You can represent the biology lab as an abstracted "mind" that can think, have intentions, and make plans (Waytz & Young, 2012; Knobe & Prinz, 2008; Strohminger & Jordan, 2022), or you can represent lab members' mental states in terms of their institutional roles (Jara-Ettinger & Dunham, 2024). Another way to simplify this process is to constrain whom you reason about, and how deeply to think. You can structure the biology lab into sub-teams to limit the number of collaborators you need to represent (Gershman & Cikara, 2023), or selectively engage in recursive reasoning only when necessary (Nasvytis et al., 2025; Charpentier et al., 2020; Wu et al., 2022). Each of these strategies reduces cognitive load, but at the cost of potentially overlooking important information within the group.

To understand how individuals shift between these shortcuts and the more costly recursive reasoning in large-scale collaboration, new methodological approaches are required. Currently, it is difficult to create large-scale social groups in traditional laboratory experiments, or test theories of large-scale social phenomena against real-world data. Recent advances have

introduced tools to study collaborations at an unprecedented scale, allowing researchers to explore how humans adapt to work together effectively in larger groups. In one study, we analyzed data from more than 20,000 players in One Hour One Life—a multiplayer online game where players build communities and develop technologies (Vélez et al., 2024). We found that the composition of communities was closely tied to their growth and decline, and that communities that balanced between maintaining and developing technologies tended to thrive. These findings open the door to understanding the strengths and weaknesses of different team structures, and using new approaches such as multiplayer online games as testbeds for theories of large-scale social dynamics.

The challenges of scaling collaboration are not limited to human teams. They also arise in human-AI and multi-agent artificial systems, where coordination must occur among agents with varying capabilities and access to information. The Collaborative Utility Calculus offers a benchmark for evaluating whether these systems behave like human teams: it predicts how agents with different capabilities should collaborate and distribute tasks based on inferred utility. This would allow for testing whether multi-agent artificial systems truly replicate the principles underlying human collaboration (Mieczkowski et al., 2025a; Mieczkowski et al., 2025b)—or merely reproduce superficial patterns (e.g., Xiang et al., 2025c; Bigelow et al., 2025). Conversely, multi-agent systems also offer a powerful theoretical tool for cognitive science. By engineering artificial agents with systematically varied cognitive constraints—such as limited planning horizons or restricted access to others' internal states—we can simulate the downstream consequences of individual limitations on group behavior. Multi-agent systems also allow us to examine the extent to which large-scale collaboration can emerge purely from simple task constraints (as in many multi-agent reinforcement learning systems), versus requiring the rich social reasoning instantiated in the Collaborative Utility Calculus. In this way, multi-agent systems are not only beneficiaries of cognitive theories, but also experimental platforms for testing and refining them.

7. Conclusion

Collaboration is a fundamental part of human life. Efforts to understand how humans collaborate have spanned decades and disciplines. Philosophical theories define the essential features that distinguish individual actions from truly collaborative activities. Evolutionary theories illuminate how collaboration can emerge and persist as a stable population-level adaptation to enhance group survival and success. Economic theories formalize the incentive structures that enable cooperation among self-interested agents. Existing psychological theories explore the cognitive mechanisms that allow individuals to infer how individuals think and act. Each of these literatures has made important contributions to a different piece of collaboration. What remains missing is a framework that connects these pieces.

Here, we propose the *Collaborative Utility Calculus,* a framework that explains how reasoning about other minds enables humans to collaborate. According to this framework, humans have an intuitive theory of competence and effort, which they knit together with their intuitive theory of belief-desire reasoning to make joint inferences and plan joint actions. We reviewed the

empirical evidence supporting this framework throughout the cognitive sciences and discussed its broad implications for cognitive science, particularly in understanding how the cognitive capacities that support collaboration emerge over development, how they shape pedagogy and communication, and how they may scale to support large-scale collaborations in both human and artificial systems.

Our framework rests on the premise that people have some information about their collaborators' competence and mental states. When this assumption holds, we have found that people are capable of making flexible inferences about others' competence and effort from minimal observations. Yet, when it does not hold, the Collaborative Utility Calculus may also be applied to understand how and why collaborations break down. It remains an open question how people acquire this information when it is impossible to observe others' contributions directly, as in cross-continental scientific collaborations, or when this information may be less reliable, as when collaborators lie about their ability or effort or strategically mislead others' inferences (Xiang et al., 2026). Under these conditions of ambiguity, people may also use simpler heuristics instead (Golman et al., 2020)—such as assuming that more confident individuals are more competent, or relying on past roles as a proxy for current ability—which may circumvent the need for directly inferring these variables and require fewer computational resources. The Collaborative Utility Calculus can serve as a benchmark to understand when certain heuristics are appropriate and what they may miss in more natural, ambiguous settings—for example, when changes in a person's competence are difficult to observe directly.

Rather than looking at these inferences at a single timepoint, another fruitful direction for future work is to consider how they evolve over the course of real-time social interaction. In dynamic collaborations, people often signal their intentions through action—by taking paths that are most direct to their goals (Baker et al., 2009) or unambiguous in context (Cheng et al., 2022). Observers use these cues, along with environmental constraints, to infer what others are capable of doing and what they are trying to do (Wu et al., 2021). Using these cues, people can coordinate effectively even without explicit communication: In real-time collaborative games, individuals converge on joint goals by tracking one another's actions (Tang et al., 2022; Candidi et al., 2015) and can predict future locations even when partners briefly disappear (Yin et al., 2016). Over time, partners can self-organize into stable, complementary roles, even when they cannot directly observe each other's actions (Goldstone et al., 2024). Modeling these interactions through the lens of the Collaborative Utility Calculus may help reveal how beliefs about others' competence and effort are updated iteratively, as people observe and act with one another in real time.

The Collaborative Utility Calculus focuses on the cognitive mechanisms that support collaboration at the individual level—how people represent their own and other people's competence, effort, and mental states to plan joint actions. Complementary traditions in psychological game theory highlight several principles that facilitate coordination at the group level. These include *common knowledge*, a shared understanding of what all parties know (Thomas et al., 2014); *salience*, the use of unique "focal points" to guide convergence on a single solution without explicit communication (Schelling, 1960); and *group identification*, the

tendency to treat the group's success as a shared payoff (Colman & Gold, 2018). Integrating these principles with the Collaborative Utility Calculus may yield a richer account of how humans navigate the challenges of collaboration under cognitive constraints. For example, common knowledge may help prevent runaway recursive reasoning about effort, while salience can reduce coordination failures by allowing collaborators to align quickly on the same plan.

Collaboration has long been seen as a hallmark of human intelligence—what separates us from other species and enables achievements that no individual could accomplish alone (Henrich, 2015). Today, collaboration is changing rapidly: People now work in large teams on a global scale, and artificial systems have the potential to transform how we work together. As the ways we collaborate evolve, so must our theories. The Collaborative Utility Calculus opens up new avenues for understanding the cognitive tools we have at our disposal to navigate this shifting landscape, and how we can better tailor collaborations to the strengths and limitations of the human mind.

Acknowledgments

We thank Fiery Cushman for helpful discussions. This work was supported by a dissertation completion fellowship to YX from Harvard University and a Templeton World Charity Foundation grant (#20648) to NV.

Funding Statement

This work was supported by a dissertation completion fellowship to YX from Harvard University and a Templeton World Charity Foundation grant (#20648) to NV.

Conflict of Interest Statement

The authors declare no conflicts of interest.

References

Bratman, M. E. (1992). Shared cooperative activity. *The Philosophical Review*, 101(2), 327-341.
Gilbert, M. (2009). Shared intention and personal intentions. *Philosophical Studies*, *144*, 167-187.
Grosz, B. J., & Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, 86(2), 269-357.
Gilbert, M. (1987). Modelling collective belief. *Synthese*, *73*(1), 185-204.
Tuomela, R. (1992). Group beliefs. *Synthese*, *91*(3), 285-318.
Butterfill, S. (2012). Joint action and development. *The Philosophical Quarterly*, *62*(246), 23-47.
Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in cognitive sciences*, *10*(2), 70-76.
Knoblich, G., Butterfill, S., & Sebanz, N. (2011). Psychological research on joint action: theory and data. *Psychology of learning and motivation*, *54*, 59-101.
Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks, 23*(8–9), 998–1003.
Pesquita, A., Whitwell, R. L., & Enns, J. T. (2018). Predictive joint-action model: A hierarchical predictive approach to human cooperation. *Psychonomic Bulletin & Review*, *25*(5), 1751-1769.
Sebanz, N., & Knoblich, G. (2021). Progress in joint-action research. *Current Directions in Psychological Science*, *30*(2), 138-143.
Candidi, M., Curioni, A., Donnarumma, F., Sacheli, L. M., & Pezzulo, G. (2015). Interactional leader–follower sensorimotor communication strategies during repetitive joint actions. *Journal of the Royal Society Interface, 12*(110).
Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: just like one's own?. *Cognition*, *88*(3), B11-B21.
Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and Brain Sciences*, 1(4), 515-526.
Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113(3), 329-349.
Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4), 0064.
von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton University Press.
Nash Jr, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, *36*(1), 48-49.
Pollack, M. E. (1990). Plans as complex mental attitudes. In P.R. Cohen, J. Morgan, M.E. Pollack (Eds.), *Intentions in Communication* (pp. 77–103). MIT Press.
Searle, J. R. (1990). Collective intentions and actions. In P.R. Cohen, J. Morgan, M.E. Pollack (Eds.), *Intentions in Communication* (pp. 401–415). MIT Press.
Camerer, C. F. (2003). *Behavioural game theory: Experiments in strategic interaction.* Russell Sage Foundation.
Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge University Press.
Bacharach, M. (2006). *Beyond individual choice: Teams and frames in game theory*. Princeton University Press.
Colman, A. M., & Gold, N. (2018). Team reasoning: Solving the puzzle of coordination. *Psychonomic Bulletin & Review*, *25*(5), 1770-1783.
Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980-994.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and human behavior*, *25*(2), 63-87.
Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology*, *65*(1), 17-28.
Török, G., Pomiechowska, B., Csibra, G., & Sebanz, N. (2019). Rationality in joint action: Maximizing coefficiency in coordination. *Psychological Science*, *30*(6), 930-941.
Török, G., Stanciu, O., Sebanz, N., & Csibra, G. (2021). Computing joint action costs: Co-actors minimize the aggregate individual costs in an action sequence. *Open Mind*, *5*, 100-112.
Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817-868.
Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *91*(1), 166-193.
Gallucci, M., & Perugini, M. (2000). An experimental test of a game‐theoretical model of reciprocity. *Journal of Behavioral Decision Making*, *13*(4), 367-389.
Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817-869.
Colman, A. M. (2024). Team reasoning cannot be viewed as a payoff transformation. *Economics & Philosophy*, *40*(1), 1-11.
Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior. *Social Philosophy and Policy*, *10*(1), 69-89.
Gilbert, M. (1990). Walking together: A paradigmatic social phenomenon. *Midwest Studies in Philosophy*, *15*(1), 1-14.
Hertz, U., Köster, R., Janssen, M. A., & Leibo, J. Z. (2025). Beyond the matrix: Experimental approaches to studying cognitive agents in social-ecological systems. *Cognition*, *254*, 105993.
Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.
Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., & Kleiman‐Weiner, M. (2021). Too many cooks: Bayesian inference for coordinating multi‐agent collaboration. *Topics in Cognitive Science*, *13*(2), 414-432.
Hawkins, R. D., Berdahl, A. M., Pentland, A. S., Tenenbaum, J. B., Goodman, N. D., & Krafft, P. M. (2023). Flexible social inference facilitates targeted social learning when rewards are not observable. *Nature Human Behaviour*, *7*(10), 1767-1776.
Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589-604.
Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14-23.
Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*(9), 1233-1235.
Raihani, N. J., & Barclay, P. (2016). Exploring the trade-off between quality and fairness in human partner choice. *Royal Society Open Science*, *3*(11), 160510.
Xiang, Y., Gershman, S. J., & Gerstenberg, T. (2026). A signaling theory of self-handicapping. *Cognition*, *266*, 106288.
Jara-Ettinger, J., & Gweon, H. (2017). Minimal covariation data support future one-shot inferences about unobservable properties of novel agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 39).
Xiang, Y., Vélez, N., & Gershman, S. J. (2023a). Collaborative decision making is grounded in representations of other people's competence and effort. *Journal of Experimental Psychology: General*, *152*(6), 1565.

Lucca, K., Horton, R., & Sommerville, J. A. (2020). Infants rationally decide when and how to deploy effort. *Nature Human Behaviour*, *4*(4), 372-379.
Leonard, J. A., Lee, Y., & Schulz, L. E. (2017). Infants make more attempts to achieve a goal when they see adults persist. *Science*, *357*(6357), 1290-1294.
Leonard, J. A., Garcia, A., & Schulz, L. E. (2020). How adults' actions, outcomes, and testimony affect preschoolers' persistence. *Child Development*, *91*(4), 1254-1271.
Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., & Gershman, S. J. (2025a). People reward others based on their willingness to exert effort. *Journal of Experimental Social Psychology*, *116*, 104699.
Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General*, *145*(12), 1654.
Jara-Ettinger, J., Kim, N., Muetener, P., & Schulz, L. (2014). Running to do evil: Costs incurred by perpetrators affect moral judgment. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36).
Gerstenberg, T., Ejova, A., & Lagnado, D. (2011). Blame the skilled. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33).
Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., & Gershman, S. J. (2023b). Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition*, *241*, 105609.
Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, *106*(3), 1465-1477.
Smith, A. (1937). The wealth of nations [1776]. *New York: Modern Library.*
Xiang, Y., Vélez, N., & Gershman, S. J. (2024). Optimizing competence in the service of collaboration. *Cognitive Psychology*, *150*, 101653.
Tian, F., Gershman, S., & Xiang, Y. (2026). Training collaborators for effective division of labor. *PsyArXiv preprint.*
McCarthy, W. P., Hawkins, R. D., Wang, H., Holdaway, C., & Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. *arXiv preprint arXiv:2107.00077*.
Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081-1085.
Larson Jr, J. R. (2013). *In search of synergy in small group performance*. Psychology Press.
Almaatouq, A., Alsobay, M., Yin, M., & Watts, D. J. (2021). Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences*, *118*(36), e2101062118.
Bernstein, E., Shore, J., & Lazer, D. (2018). How intermittent breaks in interaction improve collective intelligence. *Proceedings of the National Academy of Sciences*, *115*(35), 8734-8739.
Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *science*, *330*(6004), 686-688.
Edmondson, A. C. (2012). *Teaming: How organizations learn, innovate, and compete in the knowledge economy*. John Wiley & Sons.
Hutchins, E. (1995). *Cognition in the Wild*. MIT press.
Hackman, J. R. (2011). *Collaborative intelligence: Using teams to solve hard problems*. Berrett-Koehler Publishers.
O'mahony, S., & Ferraro, F. (2007). The emergence of governance in an open source community. *Academy of Management Journal*, *50*(5), 1079-1106.
Vélez, N., Wu, C. M., Gershman, S. J., & Schulz, E. (2024). The rise and fall of technological development in virtual communities. *PsyArXiv preprint.*
Curioni, A., Voinov, P., Allritz, M., Wolf, T., Call, J., & Knoblich, G. (2022). Human adults prefer to cooperate even when it is costly. *Proceedings of the Royal Society B: Biological Sciences, 289*(1973).
Vizmathy, L., Begus, K., Knoblich, G., Gergely, G., & Curioni, A. (2024). Better together: 14-month-old infants expect agents to cooperate. *Open Mind, 8,* 1-16.
Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291-1298.

Michael, J., Sebanz, N., & Knoblich, G. (2016). Observing joint action: Coordination creates commitment. *Cognition*, *157*, 106-113.
Mascaro, O., & Csibra, G. (2022). Infants expect agents to minimize the collective cost of collaborative actions. *Scientific Reports, 12*(1), 17088.
ATLAS collaboration. (2022). A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery. *NATURE*, *607*(7917), 52-59.
Hutchins, E. (1991). The social organization of distributed cognition.
Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., & Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human factors*, *52*(2), 203-224.
Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In *Theories of group behavior* (pp. 185-208). New York, NY: Springer New York.
Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge (acting with technology)*. The MIT Press.
Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive science*, *37*(2), 255-285.
Cooke, N. J. (2015). Team cognition as interaction. *Current directions in psychological science*, *24*(6), 415-419.
Gabora, L. (2013). Cultural evolution as distributed computation. In *Handbook of Human Computation* (pp. 447-461). New York, NY: Springer New York.
Miton, H., & Jackson, J. C. (2025). Complex technology requires cultural innovations for distributing cognition. *Trends in Cognitive Sciences*.
Vélez, N., Christian, B., Hardy, M., Thompson, B. D., & Griffiths, T. L. (2023). How do humans overcome individual computational limitations by working together?. *Cognitive Science*, *47*(1), e13232.
Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1-34.
Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287-292.
Liu, S., Karakose-Akbiyik, S., Outa, J., & Kim, M. J. (2025). How physical information is used to make sense of the psychological world. *Nature Reviews Psychology*, 1-15.
Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, *96*(1), B1-B11.
Skerry, A. E., Carey, S. E., & Spelke, E. S. (2013). First-person action experience reveals sensitivity to action efficiency in prereaching infants. *Proceedings of the National Academy of Sciences*, *110*(46), 18728-18733.
Gerson, S. A., & Woodward, A. L. (2014). Learning from their own actions: The unique effect of producing actions on infants' action understanding. *Child Development*, *85*(1), 264-277.
Liu, S., Brooks, N. B., & Spelke, E. S. (2019). Origins of the concepts cause, cost, and goal in prereaching infants. *Proceedings of the National Academy of Sciences*, *116*(36), 17747-17752.
Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038-1041.
Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301-1303.
Cortes Barragan, R., & Dweck, C. S. (2014). Rethinking natural altruism: Simple reciprocal interactions trigger children's benevolence. *Proceedings of the National Academy of Sciences*, *111*(48), 17071-17074.
Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, *10*(1), 121-125.
Warneken, F., Chen, F., & Tomasello, M. (2006). Cooperative activities in young children and chimpanzees. *Child Development*, *77*(3), 640-663.

Henderson, A. M., & Woodward, A. L. (2011). “Let’s work together”: What do infants understand about collaborative goals?. *Cognition*, *121*(1), 12-21.
Warneken, F., Steinwender, J., Hamann, K., & Tomasello, M. (2014). Young children's planning in a collaborative problem-solving task. *Cognitive Development*, *31*, 48-58.
Brownell, C. A., Ramani, G. B., & Zerwas, S. (2006). Becoming a social partner with peers: Cooperation and social understanding in one-and two-year-olds. *Child Development*, *77*(4), 803-821.
Leonard, J. A., Bennett-Pierre, G., & Gweon, H. (2019). Who is better? Preschoolers infer relative competence based on efficiency of process and quality of outcome. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 41).
Magid, R. W., DePascale, M., & Schulz, L. E. (2018). Four- and 5-year-olds infer differences in relative ability and appropriately allocate roles to achieve cooperative, competitive, and prosocial goals. *Open Mind*, *2*(2), 72-85.
Baer, C., & Odic, D. (2022). Mini managers: Children strategically divide cognitive labor among collaborators, but with a self-serving bias. *Child Development*, *93*(2), 437-450.
Dreyer, T., Haluts, A., Korman, A., Gov, N., Fonio, E., & Feinerman, O. (2025). Comparing cooperative geometric puzzle solving in ants versus humans. *Proceedings of the National Academy of Sciences*, *122*(1), e2414274121.
Seeley, T. D., Camazine, S., & Sneyd, J. (1991). Collective decision-making in honey bees: how colonies choose among nectar sources. *Behavioral Ecology and Sociobiology*, *28*(4), 277-290.
Alcalá, L., Rogoff, B., & López Fraire, A. (2018). Sophisticated collaboration is common among Mexican-heritage US children. *Proceedings of the National Academy of Sciences*, *115*(45), 11377-11384.
Dittmann, A. G., Stephens, N. M., & Townsend, S. S. (2020). Achievement is not class-neutral: Working together benefits people from working-class contexts. *Journal of Personality and Social Psychology*, *119*(3), 517.
Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896-910.
Wang, P., Wang, J., Paranamana, P., & Shafto, P. (2020). A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems*, *33*, 17582-17593.
Shafto, P., Wang, J., & Wang, P. (2021). Cooperative communication as belief transport. *Trends in Cognitive Sciences*, *25*(10), 826-828.
Chen, A. M., Palacci, A., Vélez, N., Hawkins, R. D., & Gershman, S. J. (2024). A hierarchical Bayesian model of adaptive teaching. *Cognitive Science*, *48*(7), e13477.
Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55-89.
Vélez, N., Chen, A. M., Burke, T., Cushman, F. A., & Gershman, S. J. (2023). Teachers recruit mentalizing regions to represent learners’ beliefs. *Proceedings of the National Academy of Sciences*, *120*(22), e2215015120.
Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322-330.
Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children’s reasoning about others’ knowledge and intent. *Developmental Science*, *15*(3), 436-447.
Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: how children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, *120*(4), 779.
Bass, I., Espinoza, C., Bonawitz, E., & Ullman, T. D. (2024). Teaching without thinking: Negative evaluations of rote pedagogy. *Cognitive Science*, *48*(6), e13470.
Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2021). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*, *150*(11), 2246.
Vélez, N., & Gweon, H. (2019). Integrating incomplete information with imperfect advice. *Topics in Cognitive Science*, *11*(2), 299-315.

Sumers, T. R., Ho, M. K., Hawkins, R. D., & Griffiths, T. L. (2023). Show or tell? Exploring when (and why) teaching with language outperforms demonstration. *Cognition*, *232*, 105326.
Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818-829.
Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*(1), 519-540.
Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, *4*, 71-87.
Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological Review*, *126*(3), 395.
Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002-12007.
Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, *15*(10), 694-698.
VanderBorght, M., & Jaswal, V. K. (2009). Who knows best? Preschoolers sometimes prefer child informants over adult informants. *Infant and Child Development: An International Journal of Research and Practice*, *18*(1), 61-71.
Lutz, D. J., & Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child Development*, *73*(4), 1073-1084.
Keil, F. C., Stein, C., Webb, L., Billings, V. D., & Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. *Cognitive science*, *32*(2), 259-300.
Xiang, Y., & Gershman, S. (2025b). Modeling intrinsic motivation as reflective planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 47).
Ham, H., Zhao, B., Griffiths, T. L., & Vélez, N. (2025). Teaching Recombinable Motifs Through Simple Examples. *Cognitive Science*, *49*(8), e70103.
Waytz, A., & Young, L. (2012). The group-member mind trade-off: Attributing mind to groups versus group members. *Psychological Science*, *23*(1), 77-85.
Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, *7*(1), 67-83.
Strohminger, N., & Jordan, M. R. (2022). Corporate insecthood. *Cognition*, *224*, 105068.
Jara-Ettinger, J., & Dunham, Y. (2024). The institutional stance. *Behavioral and Brain Sciences*, 1-62.
Gershman, S. J., & Cikara, M. (2023). Structure learning principles of stereotype change. *Psychonomic Bulletin & Review*, *30*(4), 1273-1293.
Nasvytis, L., Gershman, S. J., & Cushman, F. (2025). Belief neglect: A heuristic in social reasoning. *PsyArXiv preprint.*
Charpentier, C. J., Iigaya, K., & O'Doherty, J. P. (2020). A neuro-computational account of arbitration between choice imitation and goal emulation during human observational learning. *Neuron*, *106*(4), 687-699.
Wu, C. M., Vélez, N., & Cushman, F. A. (2022). Representational exchange in human social learning: Balancing efficiency and flexibility. In I. C. Dezza, E. Schulz, & C. M. Wu (Eds.), *The Drive for Knowledge: The Science of Human Information Seeking*. Cambridge University Press.
Mieczkowski, E., Turner, C., Vélez, N., & Griffiths, T. L. (2025a). People evaluate idle collaborators based on their impact on task efficiency. *Cognition*, *264*, 106200.
Mieczkowski, E., Mon-Williams, R., Bramley, N., Lucas, C. G., Velez, N., & Griffiths, T. L. (2025b). Predicting multi-agent specialization via task parallelizability. *arXiv preprint arXiv:2503.15703*.
Xiang, Y., Bigelow, E., Gerstenberg, T., Ullman, T., & Gershman, S. J. (2025c). Language models assign responsibility based on actual rather than counterfactual contributions. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 47).

Bigelow, E., Xiang, Y., Gerstenberg, T., Ullman, T., & Gershman, S. J. (2025). Actual or counterfactual? Asymmetric responsibility attributions in language models. *NeurIPS Workshop on Interpreting Cognition in Deep Learning Models.*

Golman, R., Bhatia, S., & Kane, P. B. (2020). The dual accumulator model of strategic deliberation and decision making. *Psychological Review*, *127*(4), 477.

Cheng, S., Zhao, M., Zhu, J., Zhou, J., Shen, M., & Gao, T. (2022). Intentional commitment through an internalized theory of mind: Acting in the eyes of an imagined observer. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44).

Tang, N., Gong, S., Zhao, M., Gu, C., Zhou, J., Shen, M., & Gao, T. (2022). Exploring an imagined "we" in human collective hunting: Joint commitment within shared intentionality. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44).

Yin, J., Xu, H., Ding, X., Liang, J., Shui, R., & Shen, M. (2016). Social constraints from an observer's perspective: Coordinated actions make an agent's position more predictable. *Cognition*, *151*, 10-17.

Goldstone, R. L., Andrade‐Lotero, E. J., Hawkins, R. D., & Roberts, M. E. (2024). The emergence of specialized roles within groups. *Topics in Cognitive Science*, *16*(2), 257-281.

Thomas, K. A., DeScioli, P., Haque, O. S., & Pinker, S. (2014). The psychology of coordination and common knowledge. *Journal of Personality and Social Psychology*, *107*(4), 657.

Schelling, T. C. (1960). *The strategy of conflict.* Harvard University Press.

Henrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.