TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 11 (2019) 374–392

ISSN: 1756-8765 online
DOI: 10.1111/tops.12341

This article is part of the topic "Computational Approaches to Social Cognition," Samuel Gershman and Fiery Cushman (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview

A Simple Computational Theory of General Collective Intelligence

Peter M. Krafft

Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology

Received 19 July 2017; received in revised form 16 December 2017; accepted 4 January 2018

Abstract

Researchers have recently demonstrated that group performance across tasks tends to be correlated, motivating the use of a single metric for the general collective intelligence of groups akin to general intelligence metrics for individuals. High general collective intelligence is achieved when a group performs well across a wide variety of tasks. A number of factors have been shown to be predictive of general collective intelligence, but there is sparse formal theory explaining the presence of correlations across tasks, betraying a fundamental gap in our understanding of what general collective intelligence is measuring. Here, we formally argue that general collective intelligence arises from groups achieving commitment to group goals, accurate shared beliefs, and coordinated actions. We then argue for the existence of generic mechanisms that help groups achieve these cognitive alignment conditions. The presence or absence of such mechanisms can potentially explain observed correlations in group performance across tasks. Under our view, general collective intelligence can be conceived as measuring group performance on classes of tasks that have particular combinations of cognitive alignment requirements.

Keywords: Collective intelligence; Multiagent systems; Collective agency; Collective rationality; Computational social science; Computational theory

Correspondence should be sent to Peter M. Krafft, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. E-mail: pkrafft@csail.mit.edu

1. Introduction

In a 1963 cinematic retelling of the ancient myth of Jason and the Argonauts, our heroes are faced with trial after trial in which these companions must come together as a team to overcome new difficulties—from navigating the sea together in their ship the Argo to battling a colossus to trapping an aerie of harpies. Such displays of agility in teamwork are just as present in the legends of the modern world. One popular way of organizing software engineering teams is called Scrum. The premise of Scrum is achieving tight coordination through frequent brief communication in order to readily adapt to unanticipated challenges and shifts in demands. In sports, teams are constantly faced with unique situations, some of which determine the fate of a game or a title. In the 2017 Super Bowl, the New England Patriots made the largest Super Bowl comeback in history, scoring 25 points in the final minute of the hour-and-a-half long game. This comeback allowed the team to win the title with a score of 34 points to 28 points after an overtime round. Notable displays of teamwork are not limited to highly trained or practiced groups. Researchers in crowdsourcing and human computation have repeatedly demonstrated the ability of ad-hoc human groups to accomplish tasks they have no prior experience executing. Successful examples range from scaffolded collaboration in crowdsourcing applications such as robotic control (Lasecki, Murray, White, Miller, & Bigham, 2011), scientific research (Vaish et al., 2017), or animation (Lasecki et al., 2015) to fairly open-ended collaboration in stylized games such as multiagent tracking (Krafft, Hawkins, Pentland, Goodman, & Tenenbaum, 2015) and real applications like on-the-fly disaster response crisis mapping (Mao, Mason, Suri, & Watts, 2016).

Other teams are legendary for their failures. An investigation of the 1986 Challenger disaster revealed that the spaceship's failure could have been avoidable—individual actors knew about the potential faultiness of the spaceship part that ultimately caused the disaster but, in violation of NASA regulation, did not communicate this knowledge in time. In more recent years there have been two notable cases of Internet sleuths in the related online communities reddit and 4chan identifying incorrect suspects in the 2013 Boston Marathon Bombing and the 2017 Unite the Right car attack. In both cases, the misidentification led to coordinated harassment of innocent people.

"General collective intelligence" has been defined as a summary of the performance of a group or a team across a set of heterogeneous tasks (Woolley, Aggarwal, & Malone, 2015a; Woolley, Chabris, Pentland, Hashmi, & Malone, 2010). Research on this topic has established that the performance of a team tends to be correlated across tasks (Engel, Woolley, Jing, Chabris, & Malone, 2014; Engel et al., 2015; Woolley et al., 2010). Teams that score highly on one task in these researchers' battery of tests tended to also score highly on other tasks. This finding motivated what these researchers called the "c"-factor—a single number, analogous to psychometrician's "g"-factor, defining a team's general collective intelligence and accounting for much of the variance in performance across teams and tasks.

What enables some teams to be successful across wide varieties of tasks? Why does team performance tend to be correlated across tasks? There are a number of factors that

recent studies have shown to be predictive of such capacities. Woolley et al. (2010) argued that individual team member ability was not the most important aspect in determining a team's general collective intelligence. Team members having high emotional intelligence and well-developed theory of mind is thought to be important—and the more people in the group with these traits, the better (Engel et al., 2014; Woolley et al., 2010). Group members contributing freely and equally to discussion has also been demonstrated as important (Kim, Chang, Holland, & Pentland, 2008; Woolley et al., 2010). Other work has argued that having rotating leadership can be beneficial (Gloor, Almozlino, Inbar, Lo, & Provost, 2014).

What principles lie behind these factors? There is sparse theory to unify these empirical results. Why is it the case that a single leader with high individual ability is not enough? Why does not a team just need one or two people with high emotional intelligence? Are there other ways groups can achieve general collective intelligence without these factors or are these factors necessary? We approach these questions in the present work by taking a Marrian approach. Marr (1982) innovated a framework for analyzing complex information processing systems to provide answers to "how" and "why" questions about these systems. Marr introduced a functionalist argument that information processing systems are oriented toward solving particular computational tasks using particular implementations of specific algorithms. Understanding these three aspects—identifying the computational, algorithmic, and implementation levels at play in a particular case—form the content of a Marrian analysis. The computational level provides answers to "why" questions with a teleological final or ultimate cause in the form of the fundamental computational problem being addressed by the system. The algorithmic and implementation levels provide answers to "how" questions. Previous authors, most notably Hutchins (1995), have argued that team behavior can be framed through a Marrian lens as distributed information processing. Building on this tradition, we attempt to analyze the abstract phenomenon of general collective intelligence as a computational property of varied team behavior.

In order to conduct a Marrian analysis, we define general collective intelligence as high performance of a group over a set of mathematically defined tasks. This formal definition shares the spirit of the empirical definition of general collective intelligence in capturing high performance across a variety of benchmark tasks. Our analysis then leads us to draw upon a simple result from the multiagent systems literature (Shoham & Leyton-Brown, 2008) as a starting point for a formal theory of general collective intelligence. The simple, intuitive result we present is that distributed multiagent systems act as if controlled by a rational centralized single-agent controller when agents have identical accurate beliefs, identical utility functions, and an ability to break symmetries of pure coordination problems. These alignment conditions are therefore sufficient to achieve general collective intelligence on a wide variety of mathematically defined multiagent tasks. In fact, beyond being sufficient, these conditions are actually necessary to achieve optimal performance across all tasks. Therefore, the only way to achieve ideal general collective intelligence is through precise cognitive alignment on beliefs, goals, and actions. At the same time, the strict alignment conditions of this result divorce it from practical

situations, so we proceed to make further theoretical progress using this result as a point of inspiration and an organizing principle in less constrained settings. We explore relaxations when agents have only approximately aligned beliefs, utility functions, and actions. We show that in a restricted class of tasks, approximate cognitive alignment of these attributes is still important to achieving good performance across tasks.

The key insight from this analysis is that correlations in performance across tasks exist because there are conditions that must be met for good performance on certain classes of tasks. If a group has a generic mechanism to achieve the condition necessary for that class of tasks, then the group has the capacity to be successful across the range of tasks in that class. This framework therefore predicts correlations between tasks but also that performance across tasks should not be uniformly correlated. Certain generic team abilities lead to success on certain types of tasks. Having conducted this computational analysis, we then turn to an algorithmic analysis. In this second part, we discuss generic mechanisms displayed in human group behavior that appear to partially function as if to aid in achieving these conditions for general collective intelligence. This algorithmic analysis ties together the existing factors that have been shown experimentally to promote general collective intelligence with commonly observed psychological and sociological processes that promote cognitive alignment, such as homophily, social preferences, and social learning.

2. General versus specific collective intelligence

Two separate conceptions of collective intelligence are sometimes conflated. Studies of collective intelligence are often motivated by the remarkable success certain groups have had at specific tasks. Financial markets are cited as a demonstration of the power of the wisdom of crowds to aggregate information into evaluative judgments (Hayek, 1945). Wikipedia is a fantastic display of large-scale collective intelligence in the Internet era (Benkler, Shaw, & Hill, 2015). In the literature on collective animal behavior, colonies of ants or bees use apparently exceedingly simple mechanisms to implement collective search for new homes or to help each other forage for food (Gordon, 2010; Pratt, Mallon, Sumpter, & Franks, 2002; Seeley, 1989). Surely, these examples are all rightfully called examples of collective intelligence. However, these examples are also meaningfully distinct from the phenomenon of a general factor explaining collective intelligence across a variety of tasks that has been identified by researchers of human group behavior.

Before proceeding, we must distinguish between what we will call "specific collective intelligence" and what we will call "general collective intelligence." We will call the ability of a group to accomplish a single task specific collective intelligence. We will reserve the phrase "general collective intelligence" for a measure of the performance of a group across a variety of tasks, as in the Woolley-Malone definition. The existence of general collective intelligence is not logically necessary, but empirical work has demonstrated the existence of suitable, effective metrics for general collective intelligence. In

contrast, specific collective intelligence is self-evident in any successful goal-oriented group behavior.

That said, we must also acknowledge there is not always a clear distinction between specific and general collective intelligence. The self-organization of financial markets, Wikipedia, and house-hunting, or foraging insect colonies are all examples of groups or communities successfully accomplishing specific goals. Yet these examples also include aspects of general collective intelligence. Accomplishing a complex enough task, such as compiling all of human knowledge, requires a variety of subtasks with varied requirement and challenges. Certain types of articles on Wikipedia may require different structures of collaborative organization. For instance, perhaps breaking news requires aggregation of many recent personal experiences and emerging reports, while an article on a mathematical theorem may require the dedicated attention of an individual expert. The ability of the community to flexibly adapt to these differing situations could be viewed as demonstrating a capacity somewhere in between specific and general collective intelligence.

The nuances of this issue are worth exploring more fully, but for now we will take a pragmatic approach of defining general and specific collective intelligence formally within the confines of the mathematical framework we will use. Within this framework, specific collective intelligence will refer to performance on one mathematically defined task, while general collective intelligence will refer to performance across a class of tasks. A more complete treatment of this distinction would involve taking complex cases of what we here intuitively call specific collective intelligence, such as financial markets or Wikipedia, and analyzing the variety of computational problems at play in these cases. A more complete theory that can directly accommodate these complex examples would then map the computational problems faced in these cases directly into the formalism of multiagent planning. Such an exercise could illuminate the variety and qualities of goals and subtasks involved in these cases, and thereby potentially allow us to quantify the extent to which an example is specific or general. For now, we deal only with the limiting cases of each extreme—one task or all tasks in a class of tasks.

This restriction implies that the theory we outline is most relevant to the traditional scope of laboratory studies of teamwork, wherein teams are presented either with discrete —often artificial or stylized—tasks, or with a variety of unanticipated tasks. Real examples that our theory will directly inform are well-defined highly isolated tasks like brainstorming, meeting scheduling, coordinated physical action such as following the steps of a dance. Specific collective intelligence will then refer to performance on each of these examples individually, while general collective intelligence will refer to aggregated performance across clusters of these tasks.

As an important final note drawing upon this distinction, we must remark that the focus of our work is on a theory of general collective intelligence and not on a theory of specific collective intelligence. In addition to there being commonly cited examples of the successes of collective intelligence, there are also commonly cited failures. Even while financial markets sometimes aggregate information into prices effectively, there have still been financial bubbles that have shaken society. On the more mundane scale to which our theory is more directly applicable, notable research in social psychology has

explored why some groups perform poorly at individual tasks such as group brainstorming (Nijstad & Stroebe, 2006). The theory we introduce here will not attempt to explain why group perform well or poorly on specific individual tasks. We instead offer an explanation of correlation in performance across tasks and, more precisely, why certain groups are able to achieve high performance across many varied tasks. Indeed, we will assume that the agents in our theory are endowed with the individual abilities necessary to execute whatever task is at hand, and we will focus on the interpersonal conditions that must be met for these individuals to achieve success collaboratively.

3. Conditions for general collective intelligence

We begin our theoretical treatment with a highly idealized notion of general collective intelligence, for which it is straightforward to make precise and complete theoretical statements. These idealized results illustrate the basic intuition our theory is meant to formalize. We then consider a weaker notion of general collective intelligence to demonstrate that our idealized results can be extended to more realistic settings.

3.1. Ideal general collective intelligence

The first theoretical result we present—an articulation of a simple folk theorem from the multiagent systems literature—is that rational collective agency arises from exactly aligned utilities, beliefs, and actions among individual agents. That is, there are three necessary and sufficient conditions for ideal general collective intelligence, defined as optimal performance across all of a broad class of tasks; these conditions are a group having the abilities to (a) coordinate actions, (b) align group member beliefs, and (c) align group members' utility functions with the group's shared goal. We begin by motivating this theoretical result with a selection of examples. We then offer a formal statement and an informal proof sketch of this proposition before extending it to weaker settings.

3.1.1. Motivating examples

To organize our motivating examples, the examples are grouped into classes of tasks that are directly related to each of the three conditions for general collective intelligence that we will subsequently introduce. In each case, we will highlight relevant aspects of the examples. We focus on simple examples in which a group must make just one joint action or a short series of joint actions to accomplish a well-defined collective goal since these types of simple examples are sufficient for our argument and are the types of tasks our theory directly informs.

3.1.1.1. Coordination-sensitive tasks: The first set of examples we discuss we call "coordination-sensitive" tasks. Coordination-sensitive tasks include cases where everyone needs to do the same thing or everyone needs to do different things, but it does not matter who

does what or what exactly is chosen as a consensus. For instance, in one of the standard tasks used in a Woolley-Malone collective intelligence battery, a team must coordinate to copy text from an image into a document. Supposing that the team decides to split the text up by paragraphs, it does not matter who chooses which paragraph, but it should be the case that no paragraphs are duplicated in the team. In a team consisting of Alice and Bob, one optimal joint action is for Alice to take the first paragraph and for Bob to take the second paragraph. An equally optimal joint action is for Alice to take the second and Bob the first. As another example, consider the need to listen to each other in a conversation, such as in brainstorming. To be able to hear each other, only one person can talk at a time, but there is no natural ordering of speakers. Conversational turn-taking is an important skill and a critical coordination mechanism that facilitates this process. These situations have symmetries in determining the best joint action for the group that must be broken arbitrarily.

3.1.1.2. Belief-sensitive tasks: It is also instructive to consider what we will call "belief-sensitive" tasks. Such situations have multiple good candidate joint actions a group could take like in coordination-sensitive tasks, but in belief-sensitive tasks symmetries can be broken by gathering outside information rather than by making an arbitrary choice. For example, consider the internal conflict of whether to offer an acquaintance a handshake or a hug as a greeting. Suppose that in a business context, you would both be best of with a handshake, while in a casual social interaction you would both prefer to hug. In this case, to choose between these two possible joint actions, you can use contextual information to judge what is appropriate. Other belief-sensitive situations require information aggregation to make the best choices. Perhaps you like to go to the bar after work to have a drink, and you have a group of people you enjoy seeing at the bar. All of you like going to bars with a good Old Fashioned cocktail, but you do not explicitly coordinate where you visit since your top priority is having a drink. In this case, everyone will be better off if all the friends share the information of where they have enjoyed the Old Fashioned the most. This information aggregation will then lead to you and your friends naturally congregating at the bar with the best drink.

3.1.1.3. Incentive-sensitive tasks: A third related class of situations is "incentive-sensitive" tasks. When people have very different preferences, coherent group behavior can be difficult to achieve. If people do not commit to a shared goal, and thereby implicitly agree to optimize a shared utility function, then people will end up operating independently in ways that disrupt the nominal purpose of the group. Free-riding behavior is one common occurrence of this pattern. For instance, a group of roommates might agree they want to keep their apartment clean, but if the roommates are also lazy, there is a high incentive to loaf and let the others in the group do all the work. Such behavior can lead not just to inefficiencies in accomplishing the shared goal—such as an on-average less clean apartment—but also the complete breakdown of cooperation. Another example is people not sharing their full schedules in meeting scheduling (Zou, Meir, & Parkes, 2015).

3.1.1.4. Fragile tasks: The final class of examples we consider here we call "fragile" tasks. In a fragile situation, if one person in a group makes a mistake, then everything is ruined. For instance, when two people are carrying a heavy couch up a staircase, if one person drops the couch or if the two people try to move in opposing directions too quickly, the couch will fall and someone may get hurt. Conversational turn-taking is another example of a fragile situation. If any one person talks over other people in a group discussion, the conversation can fall apart. Another interesting example is voting in a two-party system. If one party only has a slight advantage in size over the other, then the party may require unanimity in its positions to achieve its goals. Fragile tasks can be contrasted to what we will call "robust" tasks, which we will discuss more in the sections on our weaker notions of general collective intelligence.

3.1.2. Problem statement

We now move on to formulating our first theoretical result. This result, a version of a common observation in the multiagent systems literature, is that for a suitably large class of tasks, having exactly aligned beliefs, utility functions, and actions is both necessary and sufficient for optimal group performance across all tasks in that class.

We assume the setting of a finite time horizon decentralized multiagent partially observable Markov decision process (Dec-POMDP) (Bernstein, Zilberstein, & Immerman, 2000). A Dec-POMDP is a flexible representation of multiagent tasks wherein a group of agents each makes stochastic observations of the state of their environment, takes actions individually, and transitions jointly from one state to another according to stochastic dynamics determined by the current state of the environment and the set of actions taken by all the agents. More formally, such a process consists of a fixed set of N agents, discrete time, a finite time horizon T, a set of states S, a set of actions for each agent A, a distribution $P(o_i|s)$ of observations given the current environment state for each agent i, and a stochastic transition function $P(s'|s,a)$ from one state to another conditional on what actions are taken. The agents also have personal reward functions, $R_i : S \times A^N \rightarrow \mathbb{R}$, which determine the benefits each agent gets from each action in each state and which are analogous to the utility functions of those agents. In a deviation from the standard Dec-POMDP formulation, we also assume the group has a shared goal that is represented as a joint reward function $R : S \times A^N \rightarrow \mathbb{R}$. This group reward function does not directly play a role in the specified Dec-POMDP but serves as a normative standard in a mirrored Dec-POMDP that uses that group reward function for all agents, rather than those agents' individual reward functions. This final assumption restricts the theory to settings in which the group has a well-defined collaborative objective.

The best a group can achieve in this situation, in terms of the normative standard of the group reward function, is to take the joint actions that maximize the group's expected cumulative reward at each step under that reward function, according to all of the observations all group members have received up to that time: $\{\boldsymbol{a}_t^*\} = \text{argmax}_{\boldsymbol{a}_t} E[\sum_{t'=t}^{T} R(s_{t'}, \boldsymbol{a}_{t'}) | \boldsymbol{o}_{\leq t,\cdot}]$. We say that rational collective agency is achieved when the group of agents takes a joint action in this optimal set at every time step. Rational collective agency therefore corresponds to executing an optimal *centralized* multiagent POMDP

policy. The challenge in executing such a policy is that the agents are decentralized. In particular, a priori the agents only have access to their own noisy state observations and incomplete knowledge of what actions other agents will choose on a particular step. What abilities must a group possess to achieve the ideal?

It is straightforward to show that aligned utilities, beliefs, and actions are necessary and sufficient attributes for rational individual agents to have in order to achieve rational collective agency. More precisely, a group can implement an optimal centralized policy for an arbitrary Dec-POMDP if and only if (a) all agents ignore their personal reward functions, acting only to optimize the group reward function, R; (b) all agents have accurate shared beliefs, that is, each agent believes the state of the world is s with probability $P(s|o_{\leq t,\cdot})$ at each time t; and (c) agents have a coordination mechanism that allows them to choose a unique joint action from the set $\{\boldsymbol{a}_t^*\}$. Since rational collective agency is equivalent to achieving optimal performance across all tasks, these conditions are therefore necessary and sufficient for this ideal formulation of general collective intelligence.

3.1.3. Proof sketch

We now present an informal argument for why aligned utilities, beliefs, and actions are sufficient for rational collective agency. Since the agents have accurate shared beliefs, they can all compute the set of optimal actions, $\{\mathbf{a}_t^*\} = \text{argmax}_{\boldsymbol{a}_t} E[\sum_{t'=t}^{T} R(s_{t'}, \boldsymbol{a}_{t'}) | \boldsymbol{o}_{\leq t,\cdot}]$. Since the agents ignore their personal reward functions and act only to optimize R, all agents will choose an action consistent with a joint optimal action if they can. The assumption that the agents have access to a coordination mechanism ensures that a unique joint action can be selected for the group.

We can also show that these conditions are necessary for rational collective agency. If any one of the agents' utilities, beliefs, or actions are unaligned, there will exist a Dec-POMDP in which the group will fail to achieve rational collective agency. We can demonstrate this converse statement by providing examples of problems where the group fails in each case. Each problem is an example of a fragile task that is either coordination-sensitive, belief-sensitive, or incentive-sensitive. When agents act according to their own personal reward functions instead of the group reward function, we only need to postulate a one-step Dec-POMDP in which the group reward conflicts with agents' personal rewards. When agents do not have accurate aligned beliefs, we need only construct a one-step Dec-POMDP in which agents report their beliefs, and the group reward is a local proper scoring rule that incentivizes honest reporting. In this case, the group utility is maximized by all the agents reporting what would be the optimal shared beliefs—which the agents do not know and hence cannot achieve. Finally, when agents do not have a coordination mechanism, they must either attempt to choose probabilistically from the set of optimal joint actions, or choose a suboptimal joint action. Probabilistic attempts can lead to coordination failure, however. A simple N -agent pure coordination game illustrates the difficulty. In such a game the agents must each choose one of two actions, but must all choose the same action in order to receive a positive state-independent group reward. Without access to a

coordination mechanism, the agents can achieve at best a probability 0.5^N of positive reward, which is clearly a lower expected reward than the optimal joint action given the ability to coordinate.

3.2. Approximate general collective intelligence

The definitions and conditions we have given so far—that is, optimal performance across all tasks and exact alignment—are quite strict. It is not even necessarily possible to achieve these conditions given a dictatorial leader since, for example, local observations outside of the view of the leader can change individual beliefs. At the same time, there are clearly categories of tasks that do not require exact alignment. For instance, if the only tasks that were required of a group were insensitive to the beliefs of individuals, or only require individuals to be aware of their local setting, then there would be no need to align beliefs. We now consider several examples of this sort, which do not require exact alignment, to motivate extensions of our theoretical results to approximate alignment settings.

3.2.1. Motivating examples

3.2.1.1. Robust tasks: As mentioned before, one class of tasks we consider is what we call "robust" tasks, which contrast with the fragile tasks we already described. In a robust task, if at least a fraction of people perform optimally, then the group performs optimally. For instance, in a trivia game, only one person in a group needs to know the answer to a question, and if time is running short it may be better for that person to run submit the answer than to discuss with the group. Or, when carrying a couch upstairs, it is often useful for two people to have a third person spotting the couch in case one person stumbles.

3.2.1.2. Soft-margin tasks: Another similar class of tasks is what we will call "soft-margin" tasks. Unlike robust tasks, overall lack of alignment can lead to suboptimalities in soft-margin situations, but in these cases, group success is additive in the sense that if more people do better, then the group does better. For instance, in collaborative writing, if one person is a free-rider, the writing will still eventually be completed, albeit more slowly.

3.2.2. Theoretical statements

We now introduce two notions of approximate alignment based on the robust and soft-margin classes of tasks discussed in the last section. For each of these classes of problems, it is straightforward to give a simple theoretical result relating approximate alignment to group performance. In the present section, we restrict ourselves to "one-shot" games. One-shot games are a subset of Dec-PODMPs with only a single time step. To analyze the classes of robust and soft-margin tasks, we will use a notion of approximate alignment we call "n-almost-all alignment." In n-almost-all alignment, all agents except for at most n have identically aligned beliefs and utility functions, and the ability to solve pure coordination problems.

From our definition of robust tasks, it is then clear that n-almost-all alignment is sufficient for optimal joint action in robust one-shot games. As in our proof sketch for the ideal alignment case, necessity follows from counterexamples that can easily be generated by adapting the counterexamples in the ideal case and adding n players whose actions are irrelevant to the payoffs.

We now focus on a specific subclass of soft-margin tasks—those with independent additive reward structures, wherein the group utility function is the sum of individual reward functions $R = \sum_i R_i$ and the reward of each agent is independent of the other agents' actions. Furthermore, we will suppose that rewards are normalized so that $R_i \in [0, 1]$. This class of tasks captures problems that can be parallelized without complex coordination between people, such as people in different patches of a farm picking fruit. In such cases, n-almost-all alignment is sufficient to achieve at least a proportion $\frac{N-n}{N}$ of the maximum group reward. n-almost-all alignment of beliefs is also necessary to achieve at least $\frac{N-n}{N}$ performance across all tasks in this case, but alignment of utility functions and ability to solve coordination games can be arbitrarily bad without penalty since agent rewards and actions are independent.

A further relaxation can be made in each of these cases when the state space is discrete. In this case, beliefs and utility functions only need to be as aligned as to uniquely identify a finite set of best joint actions, not exactly aligned. When states that are easily confused based on observed information available lead to joint actions with similar reward, additional statements could be made. Much more could be said about approximate alignment conditions and group performance. Further work in this area could be illuminating.

3.3. Discussion

Our extensions were meant to illustrate that alignment is a theoretically important factor even for approximate general collective intelligence. While ideal general collective intelligence was straightforward to characterize completely due to the strictness of our definitions, our hodgepodge of results for approximate general collective intelligence is far from a complete theory. Nevertheless, the individual theoretical results we have motivate the investigation of alignment conditions in broader settings than the unrealistically idealized case.

One caveat should be noted. In all cases, we have supposed here that agents execute the action corresponding to their part in the best joint action they perceive. This assumption is at odds with purely rational individual action, because for agents to truly act rationally, they must have confidence in the belief states of others to justify their actions. In general, common knowledge of the beliefs, utilities functions, and shared plans is needed to guarantee individual rationality. Algorithms exist for individual agents to arrive at common knowledge and approximations of common knowledge (Krafft, Baker, Pentland, & Tenenbaum, 2016), assuming shared prior beliefs. In the present work, we omit subtleties introduced by this important requirement by supposing agents act in what they perceive to be accordance with optimal joint action, such as by participating in team reasoning (Sugden, 2015). Recent empirical work studying human coordination in simple

tasks lends plausibility to such an assumption (Kleiman-Weiner, Ho, Austerweil, Littman, & Tenenbaum, 2016).

4. Relationship with existing theories

The theoretical results we have presented shed light on the empirical findings in the literature on general collective intelligence. Theory of mind, rotating leadership, and equitable contribution may all service cognitive alignment. Theory of mind promotes inference of others' beliefs, goals, and plans. Rotating leadership and equitable contribution could be useful for both information aggregation and reaching consensus on group utility functions. Other work has argued that informal social networks in organizations are important for information flow, leading to shared beliefs that more fragmented organizational networks would not achieve (Pentland, 2014).

Our theoretical results also inform ongoing debates within the literature on general and specific collective intelligence. On the question of the role of leadership our theory suggests that a leader could be helpful is resolving pure coordination problems but may cause issues if too dominating in terms of insisting upon certain beliefs, thereby inhibiting information aggregation, or upon certain goals, thereby threatening the adoption of a shared goal. Several recent pieces of work have explored the advantages of group diversity for collective intelligence (Hong & Page, 2004, unpublished data). There are clear information aggregation benefits to having diverse sources of information to draw upon. Alignment on utility functions may be more difficult to achieve in diverse groups, but then, perhaps if an ultimate consensus utility is achieved it could be better for more people outside the group. Group size has also been a topic of interest (Kao & Couzin, 2014; Vicente-Page, P′erez-Escudero, & Polavieja, 2017). It is plausible that alignment is easier to achieve in small groups due purely to statistical effects of distributions of beliefs and preferences.

Our approach is also related to several existing frameworks. Coordination theory was an early attempt to explain collective intelligence, particularly in business settings, that focused on how organizations achieve complex coordination of activities (Malone & Crowston, 1990). The computational framework Shared-Plans offers a normative view of coordinated activities (Grosz & Kraus, 1996). We can view these prior frameworks as mainly concerned with how to achieve one of the three conditions that we emphasize, namely aligning actions. Well-known decompositions of task types, similar in spirit to the classes of task examples we discuss, are Steiner's process groupings (Steiner, 1972) and McGrath's Circumplex (McGrath, 1984). A recent review of the teamwork literature oriented toward a collective intelligence audience highlighted the importance of groups having clear goals, aligned incentives, and an ability to coordinate (Woolley, Aggarwal, & Malone, 2015b). The literature on team mental models has also stressed the importance of team members having aligned understandings of the task at hand, the role each team member plays, and in some cases the environment the team operates in (Mohammed, Ferzandi, & Hamilton, 2010).

5. Social processes that promote alignment

Given the importance of achieving the conditions for general collective intelligence, we now turn to generic mechanisms groups may have at their disposal for achieving or approaching these conditions. The generic behavioral mechanisms we will discuss have diverse meanings and effects in various contexts. For instance, one mechanism we will discuss is social learning. Social learning occurs through a broad array of behaviors. Social learning can mean being convinced by someone else's argument. Social learning can mean simply learning from observing someone else's action. In either case, social learning leads to more aligned beliefs. Social learning can also lead to negative outcomes such as social proof of false beliefs. Regardless of whether the alignment resulting from these mechanisms is an epiphenomenon, and regardless of whether there are additional negative outcomes associated with these mechanisms, our perspective is that alignment is still an additional resultant inherent positive force in group contexts in addition to these other positive and negative aspects of the mechanisms. All the mechanisms we discuss are complex, but we focus on the functional relevance of these mechanisms for understanding how groups achieve the conditions for general collective intelligence.

5.1. Aligning preferences

The first condition for rational collective agency we discuss is aligning utility functions. General collective intelligence requires commitment to a shared goal, and the utility loss that people suffer will be smaller if their own utility functions are closer to those that lead to the achievement of the group's shared goal. Multiple social processes lead to people either adopting better aligned utility functions or creating groups that have better aligned utility functions. We discuss two important social processes that lead to such alignment: homophily and social preferences.

5.1.1. Homophily

Homophily is the widely observed empirical regularity in the study of human social networks that people who are similar to each other tend to interact preferentially (McPherson, Smith-Lovin, & Cook, 2001). Homophily has been observed along a range of attributes, from various demographic characteristics (McPherson et al., 2001) to personality traits (Youyou, Stillwell, Schwartz, & Kosinski, 2017). If homophily occurs along the dimensions of people's utility functions, then this prominent social process would naturally lead to groups that have better aligned utility functions.

5.1.1.1. Example: Consider the example of finding an agreeable place for dinner with friends. Two potential situations are that everyone in a group of friends is vegetarian, or only half of the friends are vegetarian. To make the example concrete, suppose there are four friends and two restaurants—one vegetarian restaurant and one restaurant focused on meat. Suppose vegetarians obtain an individual reward of 1 for the vegetarian and 0.1 for

the meat-focused restaurant, with the flesh-eaters receiving 0.1 for the vegetarian and 1 for the other. In this case, the maximum utility loss will be zero for the homogeneous group, and 0.9 for the mixed group. General collective intelligence is also about achieving good performance on a range of tasks. What other benefits might preferentially interacting with other vegetarians have? If instead of going out to eat, the friends decide to cook dinner together, then they will enjoy each other's cooking more. Another benefit could be the conversational topics the group chooses. Perhaps the vegetarian friends enjoy talking together about the challenges or benefits of being vegetarian.

5.1.2. Social preferences

Another hypothesized feature of human behavior that leads to better aligned utilities is social preferences. Social preferences are a postulated modification to individual utility calculus that incorporates the outcomes of other actors into one's personal utility function (Gintis, 2009). Social preferences have some empirical support, and are advocated as a solution to puzzles in observed human behavior, such as excess levels of cooperation as compared to the predictions of pure individual rational action (Rabin, 2006).

5.1.2.1. Example: Take the example of a couple maintaining a home together. Suppose one partner has a preference for cleanliness while the other partner does not. Let us formalize this situation as a one-shot game in which each partner decides whether to clean or not. If nobody cleans, no reward is received by either partner. For each person that cleans, that person incurs a cost of 0.1, and the picky partner receives a reward of 0.5. When both partners commit to the shared goal of keeping their home as clean as possible, then the picky partner will receive a reward of 0.9 and the sloppy partner a reward of -0.1. Now, supposing that both partners equally weight each others' rewards in their social preference functions, then both partners will receive a reward of 0.4 for keeping the home clean together. Once again, we can also observe that social preferences are not limited in their advantage to just this single task. When deciding how warm to keep the home or deciding what household items to purchase, partners' considering each other's well-being will lead to less friction in joint decision-making.

5.2 Aligning beliefs

The second condition for rational collective agency is accurate aligned beliefs. In order for groups to make accurate decisions, information must be aggregated. Several features of human cognition and social interaction support belief agreement processes. Here, we discuss the prominent mechanism of social learning.

5.2.1. Social learning

Social learning—the ability of people to learn from observing each other—is widespread and has been hypothesized to underlie human society's remarkable ability at cultural accumulation (Boyd, Richerson, & Henrich, 2011). There is still debate about whether people engage in social learning through rational mechanisms, heuristic

mechanisms, or semi-rational mechanisms (Spyrou, 2013). Regardless, these models all share the characteristic of bringing peoples' beliefs to be closer to one another.

5.2.1.1. Example: Consider two students presenting a group project. Suppose the students are asked a question by the teacher. One student answers, and the other student must judge whether or not to correct the first student. At the same time, the second student might be unsure of her own understanding and may reason that the first student had some positive private information she does not. We can quantify this example with a simple signal processing formulation. Suppose each student received a noisy binary private signal from personal study about the yes-or-no question asked by the teacher, and suppose that the second student receives a noisier signal. In this case, even if the second student has conflicting private information, that student will infer the positive private information of the first student, and then defer to the response of the first student. In the end, the students have better aligned beliefs and a higher expected chance of avoiding embarrassment or lost points. Social learning could also be useful in this group project when the group is preparing for their presentation, for instance, to recognize and mend misunderstandings.

5.3. Aligning actions

The final condition for rational collective agency is aligned actions. Individuals must be able to select actions that mesh well together. The group must be able to select one joint action from the set of optimal joint actions. Certain components of an optimal joint action may be risky in the sense that an individual's action can be bad if others do something unexpected, and hence the individual taking a risky action must be supported by the rest of the group in the group's joint action. Here, we discuss how well-defined roles in a group context can aid in coordination.

5.3.1. Roles

By roles, we here mean the specified sets of functions assigned to each member of a team. Roles are useful for coordination in repeated interactions and critical for coordination in fast-paced situations. Roles allow team members to break the symmetries inherent in coordination problems, because they prescribe how those symmetries should be broken.

5.3.1.1. Example: For example, consider the roles of offense and defense in many types of sports teams. Setting aside the fact that roles also provide a benefit via specialization, if an American football team had to decide on-the-fly once gameplay commences how to arrange themselves, the team would be much less effective at advancing or protecting themselves from a more organized team. In a stylized formalization of this example, we can suppose that a game proceeds in two time steps. Breaking a coordination symmetry takes one time step. Attacking or defending also takes one time step. A team that has at least one attacker can break through the other team, if and only if that team has no defenders. In this case, a pre-organized team can attack the undefended unorganized team

with an offensive player on the first time step while the unorganized team is breaking their coordination symmetry.

6. Discussion

We now discuss potential extensions and limitations of our work. One question we might ask is whether our framework could also inform thinking about much larger scale social systems. There is reason to believe that it might. For instance, we can view social norms whose violations are enforced by social sanctioning as another mechanism for utility function alignment, such as norms around giving up one's seat on the bus or subway. Such norms are highly specific to this particular context, but assist in the prosocial goal of easing the worst-case experience of public transport. As another example, the convention of driving on the right or the left side of the road is a mechanism for coordination that serves the goal of safe and efficient traffic flow. The media's role of educating the public about political candidates during elections serves to align shared beliefs. These examples suggest that aligned beliefs, utilities, and actions may be important in larger scale systems than the small group contexts we have considered as our primary target of analysis.

There are other important issues that our treatment has left open. As we have noted, we are far from having a complete theory of approximate alignment and weak general collective intelligence. More work is also needed to better understand the spectrum between specific and general collective intelligence. Deeper challenges to our account include particular cases and classes of tasks where alignment may be detrimental. For example, team diversity has been positively related to collective intelligence. Our theory suggests that initial diversity can be beneficial but that eventual cognitive alignment leads to smoother group functioning. Another issue is a class of tasks that potentially falls out of the current scope of rational analysis: moral reasoning. McGrath's well-known decomposition of group tasks includes a category for moral problems. It is interesting to note that in Woolley-Malone general collective intelligence test batteries, scores across many of the tasks tend to be correlated, but performance on collective moral judgment tends to be less correlated. At the same time, most of the tasks in this battery are mainly coordination-sensitive, so perhaps this battery unintentionally tests only the aspects of general collective intelligence related to the action alignment condition. This observation suggests that an enriched test battery with belief-sensitive or incentive-sensitive tasks might more completely measure general collective intelligence, and doing so might lead to observation of less uniform correlation in performance across tasks. Classes of tasks that test aligned shared beliefs or utilities may also correlate better with moral tasks. Or perhaps not; perhaps understanding group moral judgments would require replacing the rational actor framework within our account with a richer agent model. A related concern is that motivation, team culture, identity, personality, and emotional factors are other attributes that are poorly captured by the rational actor model in our current account. Unlike moral judgment, though, these latter factors are more plausibly operating at a lower level of Marr's hierarchy than the computational level we focus on.

7. Conclusion

We have offered a new theoretical lens to understand the conditions for general collective intelligence inspired by a simple folk theorem from the multiagent systems literature. Our theoretical developments help to unify existing factors that have been offered as predictive of general collective intelligence in teams, and they provide a more nuanced perspective into what correlations in performance might be expected to be observed across tasks in teams. Our account also provides a lens for functionalist interpretations of a variety of social phenomena that result in aligned beliefs, goals, and actions.

One practical observation we have made that could lead to a more comprehensive collective intelligence battery while simultaneously validating our framework is that much work on general collective intelligence has focused on tasks that are coordination heavy. Our theory suggests that designing belief-sensitive and incentive-sensitive tasks for a new battery might better highlight other aspects of general collective intelligence that some teams capable at coordination might fail at. Our framework predicts that teams with an ability to align beliefs should be effective at belief-sensitive tasks and teams with an ability to align utility functions should be effective at incentive-sensitive tasks, but both types of teams could in principle perform poorly at coordination-sensitive tasks if they lack mechanisms for coordination. We expect that a new collective intelligence battery covering all these different classes of tasks would expose structured correlations across teams. Some teams would perform well only in tasks from just one category; other teams would perform well in a pair of categories; and some teams would perform well across all or no categories. The frequency of each pattern of correlations would depend on the frequency of the abilities to achieve each cognitive alignment condition, as well as the co-occurrence frequencies of these abilities within individuals.

Acknowledgments

This work was supported in part by the NSF GRFP under grant no. 1122374. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors. Special thanks to John Tsitsiklis, Leslie Pack Kaelbling, and Tom Malone for providing critical feedback on earlier versions of this work, and to Sandy Pentland and Josh Tenenbaum for stimulating the line of inquiry. Thanks also to our anonymous reviewers, who provided invaluable feedback that led to a vastly improved manuscript.

References

Benkler, Y., Shaw, A., & Hill, B. (2015). Peer production: A form of collective intelligence. In M. Bernstein & T. Malone (Eds.), *The handbook of collective intelligence* (pp. 175–203). Cambridge, MA: MIT Press.

Bernstein, D. S., Zilberstein, S., & Immerman, N. (2000). The complexity of decentralized control of markov decision processes. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence* (pp. 32–37). San Francisco: Morgan Kaufmann.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108*, 10918–10925.

Engel, D., Woolley, A. W., Aggarwal, I., Chabris, C. F., Takahashi, M., Nemoto, K., Kaiser, C., Kim, Y. J., & Malone, T. W. (2015). Collective intelligence in computer-mediated collaboration emerges in different contexts and cultures. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3769–3778). New York: ACM.

Engel, D., Woolley, A. W., Jing, L. X., Chabris, C. F., & Malone, T. W. (2014). Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence equally well online and face-to-face. *PLoS ONE*, *9*(12), e115212.

Gintis, H. (2009). *The bounds of reason: Game theory and the unification of the behavioral sciences*. Princeton, NJ: Princeton University Press.

Gloor, P. A., Almozlino, A., Inbar, O., Lo, W., & Provost, S. (2014). Measuring team creativity through longitudinal social signals. arXiv preprint arXiv:1407.0440.

Gordon, D. M. (2010). *Ant encounters: Interaction networks and colony behavior*. Princeton, NJ: Princeton University Press.

Grosz, B. J., & Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, *86*(2), 269–357.

Hayek, F. A. (1945). The use of knowledge in society. *American Economic Review*, *35*(4), 519–530.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, *101* (46), 16385–16389.

Hong, L., & Page, S. E. (2008). Some microfoundations of collective wisdom. In *Collective wisdom*, (pp. 56–71). San Francisco: Morgan Kaufmann.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Kao, A. B., & Couzin, I. D. (2014). Decision accuracy in complex environments is often maximized by small group sizes. In *Proceedings of the Royal Society B*, *81*(1784).

Kim, T., Chang, A., Holland, L., & Pentland, A. S. (2008). Meeting mediator: Enhancing group collaborationusing sociometric feedback. In American Economic Review, 35(4), 519-530. *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, (pp. 457–466). New York, NY: ACM.

Kleiman-Weiner, M., Ho, M., Austerweil, J., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction. In A. Papafragou et al. (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1679–1684), Austin, TX: Cognitive Science Society.

Krafft, P. M., Baker, C. L., Pentland, A. S., & Tenenbaum, J. B. (2016). Modeling human ad hoc coordination. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, (pp. 3740–3746). AAAI Press.

Krafft, P. M., Hawkins, R. X., Pentland, A., Goodman, N., & Tenenbaum, J. B. (2015). Emergent collective sensing in human groups. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings & P. P. Maglio (Eds.), *Annual conference of the cognitive science society (CogSci)*. Austin, TX: Cognitive Science Society.

Lasecki, W. S., Kim, J., Rafter, N., Sen, O., Bigham, J. P., & Bernstein, M. S. (2015). Apparition: Crowdsourced user interfaces that come to life as you sketch them. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1925–1934). New York: ACM.

Lasecki, W. S., Murray, K. I., White, S., Miller, R. C., & Bigham, J. P. (2011). Real-time crowd control of existing interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (pp. 23–32). New York: ACM.

Malone, T. W., & Crowston, K. (1990). What is coordination theory and how can it help design cooperative work systems? In F. Halasz (Ed.), *Proceedings of the 1990 ACM Conference on Computer-Supported Cooperative Work*, (pp. 357–370). New York: ACM.

Mao, A., Mason, W., Suri, S., & Watts, D. J. (2016). An experimental study of team size and performance on a complex task. *PLoS ONE*, *11*(4), e0153048.

Marr, D. (1982). *Vision: A computational approach*. San Francisco: W.H. Freeman.

McGrath, J. E. (1984). *Groups: Interaction and performance*, Vol. 14. Englewood Cliffs, NJ: Prentice-Hall.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444.

Mohammed, S., Ferzandi, L., & Hamilton, K. (2010). Metaphor no more: A 15-year review of the team mental model construct. *Journal of Management*, *36*(4), 876–910.

Nijstad, B. A., & Stroebe, W. (2006). How the group affects the mind: A cognitive model of idea generation in groups. *Personality and Social Psychology Review*, *10*(3), 186–213.

Pentland, A. (2014). *Social physics: How good ideas spread—The lessons from a new science*. New York: Penguin.

Pratt, S. C., Mallon, E. B., Sumpter, D. J., & Franks, N. R. (2002). Quorum sensing, recruitment, and collective decision-making during colony emigration by the ant leptothorax albipennis. *Behavioral Ecology and Sociobiology*, *52*(2), 117–127.

Rabin, M. (2006). The experimental study of social preferences. *Social Research: An International Quarterly*, *73*(2), 405–428.

Seeley, T. D. (1989). The honey bee colony as a superorganism. *American Scientist*, *77*(6), 546–553.

Shoham, Y., & Leyton-Brown, K. (2008). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge, UK: Cambridge University Press.

Spyrou, S. (2013). Herding in financial markets: A review of the literature. *Review of Behavioural Finance*, *5* (2), 175–194.

Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.

Sugden, R. (2015). Team reasoning and intentional cooperation for mutual benefit. *Journal of Social Ontology*, *1*(1), 143–166.

Vaish, R., Gaikwad, S. N. S., Kovacs, G., Veit, A., Krishna, R., Arrieta Ibarra, I., Simoiu, C., Wilber, M., Belongie, S., & Goel, S., et al. (2017). Crowd research: Open and scalable university laboratories. In R. Vaish et al. (Eds.), *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (pp. 829–843). New York: ACM.

Vicente-Page, J., P′erez-Escudero, A., & dePolavieja, G. G. (2017). Dynamic choices are most accurate in small groups. In *Theoretical ecology* (pp. 71–81). Dordrecht, the Netherlands: Springer.

Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015a). Collective intelligence and group performance. *Current Directions in Psychological Science*, *24*(6), 420–424.

Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015b). Collective intelligence in teams and organizations. In M. Bernstein & T. Malone (Eds.), *The handbook of collective intelligence*. Cambridge, MA: MIT Press.

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.

Youyou, W., Stillwell, D., Schwartz, H. A., & Kosinski, M. (2017). Birds of a feather do flock together: Behavior-based personality-assessment method reveals personality similarity among couples and friends. *Psychological Science*, *28*(3), 276–284.

Zou, J., Meir, R., & Parkes, D. (2015). Strategic voting behavior in Doodle polls. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 464–472). New York: ACM.