TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 17 (2025) 374–391

ISSN: 1756-8765 online
DOI: 10.1111/tops.12705

This article is part of the topic "Building the Socio-Cognitive Architecture of COHUMAIN: Collective Human-Machine Intelligence," Cleotilde Gonzalez, Henny Admoni, Scott Brown and Anita Williams Woolley (Topic Editors).

Adversarial Dynamics in Centralized Versus Decentralized Intelligent Systems

Levin Brinkmann,[a] Manuel Cebrian,[b,c] Niccolò Pescetelli[d,e]

[a]*Center for Humans and Machines, Max Planck Institute for Human Development*
[b]*Department of Statistics, Universidad Carlos III de Madrid*
[c]*UC3M-Santander Big Data Institute, Universidad Carlos III de Madrid*
[d]*Department of Humanities and Social Sciences, New Jersey Institute of Technology*
[e]*PSi, People Supported Technologies Ltd*

Received 31 May 2022; received in revised form 8 October 2023; accepted 11 October 2023

Abstract

Artificial intelligence (AI) is often used to predict human behavior, thus potentially posing limitations to individuals' and collectives' freedom to act. AI's most controversial and contested applications range from targeted advertisements to crime prevention, including the suppression of civil disorder. Scholars and civil society watchdogs are discussing the oppressive dangers of AI being used by centralized institutions, like governments or private corporations. Some suggest that AI gives asymmetrical power to governments, compared to their citizens. On the other hand, civil protests often rely on distributed networks of activists without centralized leadership or planning. Civil protests create an adversarial tension between centralized and decentralized intelligence, opening the question of how distributed human networks can collectively adapt and outperform a hostile centralized AI trying to anticipate and control their activities. This paper leverages multi-agent reinforcement learning to simulate dynamics within a human–machine hybrid society. We ask how decentralized intelligent agents can collectively adapt when competing with a centralized predictive algorithm, wherein prediction involves

Correspondence should be sent to Levin Brinkmann, Center for Humans and Machines, Max Planck Institute for Human Development, Lentzeallee 94, 14195, Berlin, Germany. E-mail: brinkmann@mpib-berlin.mpg.de

suppressing coordination. In particular, we investigate an adversarial game between a collective of individual learners and a central predictive algorithm, each trained through deep Q-learning. We compare different predictive architectures and showcase conditions in which the adversarial nature of this dynamic pushes each intelligence to increase its behavioral complexity to outperform its counterpart. We further show that a shared predictive algorithm drives decentralized agents to align their behavior. This work sheds light on the totalitarian danger posed by AI and provides evidence that decentrally organized humans can overcome its risks by developing increasingly complex coordination strategies.

Keywords: Collective intelligence; Artificial intelligence; Predictive algorithms; Decentralized intelligence; Autocurriculum; Multi-agent reinforcement learning; Coordination games

1. Introduction

Competition between machines and humans has been a recurring topic in popular culture, in movies like *"The Matrix" and "The Terminator."* Science fiction authors have long envisioned the danger of superhuman machines oppressing humans. While superhuman artificial intelligence (AI) in the real world does not yet have the autonomy and agency depicted in popular culture, AI in general is increasingly becoming part of our everyday lives. A large class of machine learning methods, commonly referred to as AI, are essentially predictive machines. Scholars have observed that when predictive AI is applied to human behavior, it generally aligns with the incentives of the AI's developers rather than the incentives of the humans whose behavior is predicted. Zuboff suggests that even in the case of aligned incentives between AI and consumers, the instrumental use of predictive algorithms does undermine human freedom of will (Zuboff, 2019).

The history of human cultural evolution has shown remarkable flexibility in humans' responses to changing and challenging environments. In this study, we leverage multi-agent reinforcement learning to simulate a human–machine hybrid society and ask how humans might collectively adapt their coordination behavior when challenged by an adversarial predictive algorithm. In particular, we investigate the coupling of decentralized intelligent agents (simulating humans) with a centralized intelligent agent (simulating AI) and the conditions under which both systems can mutually advance each other.

Humans who organize in groups show collective intelligence. The concept of collective intelligence is based on the observation that intelligent solutions can emerge from information aggregation processes among a population of simpler distributed agents (Couzin, 2007). In humans, collective intelligence can remain largely independent from the intelligence of the individual members of the group (Riedl, Kim, Gupta, Malone, & Woolley, 2021; Woolley, Chabris, Pentland, Hashmi, & Malone, 2010). Some researchers classify entities of collective intelligence by their topology of power and inner structure into hierarchies, markets, democracies, and communities (Malone, Foster, & Audio, 2018). In this work, we focus on topologies used by AI to share information during learning and inference. The high communication bandwidth of computer systems allows AI to be trained and deployed centrally. Correspondingly, we compare three predictive AI architectures in this work that differ in their centrality, that is, in their ability to share learnings across the system and to observe the full system during inference.

Crucially, AI's computation, memory, communication, and time are subject to much fewer constraints than humans (Gillings, Hilbert, & Kemp, 2016; Griffiths, 2020). Many cognitive skills are heuristics and cognitive tools to mitigate these specific human limitations (Heyes, 2018; Kahneman, 2011). Therefore, we should also expect AI, which might rely less on heuristics, to be different from human intelligence (Griffiths, 2020) and human collective intelligence. For instance, while humans' speech information rate is only $\sim$39 bits/s (Coupé, Oh, Dediu, & Pellegrino, 2019), computer systems archieve bandwidths nine magnitudes higher. In light of this communicative bottleneck, it appears unsurprising that effective group communication (social sensitivity, equal contributions of members, and the total amount of words spoken) predicts collective intelligence in humans (Woolley, Aggarwal, & Malone, 2015; Woolley et al., 2010). While the distributed nature of collective intelligence and the communication between agents is important for humans, it might be much less relevant for AI. AI systems are typically trained centrally (with some notable exceptions; Konečný et al., 2016), and all users interact with clones of the same algorithm. The large language model ChatGPT, for instance, is a single algorithm shared across many users: a central trained model queried by individual applications via the internet (Hoy, 2018). Importantly, while training is central, the model is handling each interaction with different users individually at runtime. In this work, we investigate the consequences of a human collective interacting with artificial entities that vary in the degree of centralization.

The alignment of incentives is a central question of human–AI interaction (Russell, 2010). Predictive AI could endanger some human groups if incentives are misaligned, such as when AI is controlled by authoritarian regimes or a few influential individuals (Zuboff, 2019). Collective intelligence often correlates with coordination (Kittur, Lee, & Kraut, 2009) and the ability to predict the emotional states of collaborators (Engel, Woolley, Jing, Chabris, & Malone, 2014), suggesting that behaving in a way that is predictable by peers might increase a group's collective intelligence. However, while predictability can improve human coordination, it might also make collectives vulnerable to an adversarial predictive AI. We are not arguing that AI is generally at odds with human collective intelligence. Plenty of examples exist where algorithms and humans enhance each other (Chen, Ning, Nugent, & Yu, 2020; Dellermann, Ebel, Söllner, & Leimeister, 2019; Pescetelli, 2021). For instance, algorithmically aggregating individual predictions can help groups make better collective decisions (Surowiecki, 2005). However, if incentive misalignment emerges, we argue that predictive AI could challenge human collective intelligence. In this work, we are investigating the dynamics of such competition.

A competition between AI and humans might have surprising consequences. Scientists have long drawn parallels between genetic evolution and cultural evolution as two systems that change over time by selection mechanisms (Mesoudi, 2017) and that lead to adaptation to the environment. Likewise, AI systems based on reinforcement learning can adapt their behavior to the environment. Some authors suggested a unifying framework to describe genetic, cultural, and algorithmic evolution as systems of adaptive learning units (Leibo, Hughes, Lanctot, & Graepel, 2019). Like the gene-culture co-evolution (Boyd & Richerson, 1988), this paper explores AI–human cultural co-evolution.

In multi-agent systems, the environment of each agent includes all other agents. An optimal behavior of one agent is, therefore, always contingent on the behavior of the other agents, especially if these agents are in either a competitive or cooperative relationship (Leibo et al., 2019). As a result, the behavioral change of one agent creates an exogenous challenge to all other agents. This exogenous challenge can induce adaptations of those other agents—and vice versa. This relationship can lead to an "arms race" in which all agents advance by increasing the complexity of their behavior (Dawkins & Krebs, 1979). Such a sequence of challenges has been called *autocurriculum* (Leibo et al., 2019). Autocurricula provide each learner with a dynamic environment that increases in difficulty at the same pace as the agent's increasing capabilities. The complexity shown by the system is thereby no longer limited by the external environment but by the system's adaptive capabilities, that is, the intelligence of the agents. In this work, we study an autocurriculum through human–machine competition.

In this study, we examine the dynamics of a competition between decentralized agents (humans) and a centralized agent (AI) in a multi-agent system. Our focus is on the collective dynamics of asymmetrical, interconnected learning systems where a centrally controlled, predictive agent has a topological advantage. We investigate this potential conflict in a simple strategic game. A fictitious example of such a scenario could be human protesters coordinating their actions, while algorithmic police forces aim to disrupt this coordination. In this paper, we will refer to agents collectively solving the coordination problem as collective intelligence (CI) and those that address the prediction problem as predictive intelligence (PI).

Reinforcement learning nicely captures the behavioral plasticity of both humans and algorithms (Sutton & Barto, 2018). Here, we use a multi-agent reinforcement learning approach to model the dynamics of a simple simulated human–machine society. Agent-based models have traditionally been used to investigate cultural evolution (Boyd & Richerson, 1988; Henrich & McElreath, 2003; Mesoudi, 2017). This approach has recently been applied to collective intelligence (Reia, Amado, & Fontanari, 2019). The objective of agent-based models is to investigate the emerging collective phenomena of agents following relatively simple rules (Niazi & Hussain, 2011). On the other hand, multi-agent reinforcement learning has shown an inverse phenomenon: the emergence of complex individual behavior, such as tool usage (Baker et al., 2019) or norms (Köster et al., 2022) as a result of multi-agent interactions. As adaptation is the critical property of evolution, we consider multi-agent reinforcement learning a promising approach to studying co-evolving human–machine systems.

In the literature on multi-agent reinforcement learning, either group of agents (Baker et al., 2019; Liu et al., 2019) or individual agents (Sandholm & Crites, 1996) have been pitted against each other. We deviate from this approach by letting a group of agents (the CI) compete with a single agent (the PI). We believe this nicely captures realistic settings in which a single trained AI model (e.g., BERT, GPT-3, AlexNet) can find large-scale applications for the entire human population via API calls. To model the perceived danger of an AI oppressing human coordination, we investigate an asymmetric game. The CI agents are rewarded for coordination, and the PI is rewarded for correctly predicting individual human actions. This setup loosely fits observed protests in recent history, like the Black Lives Matter (BLM)

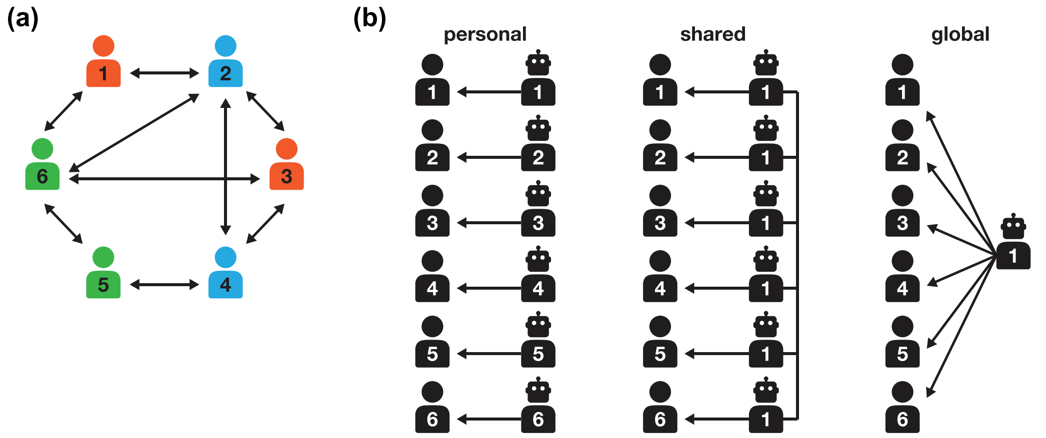

Fig. 1. (a) Agents of the collective intelligence (CI) coordinate on a network. Agents 1 and 3 receive a positive reward as their neighbors choose a different color than themselves. Agents 2, 4, 5, and 6 receive a negative reward due to a color conflict with at least one neighbor. (b) The predictive intelligence (PI) predicts the colors chosen by the CI. We investigate three different architectures. In the personal case, each member of the CI is predicted by an independent PI agent (i.e., one model per CI agent). In the shared condition, CI agents are predicted by a copy of the same PI agent (i.e., a single model predicts each individual separately). In the global case, a single PI is simultaneously predicting the colors of all CI agents (i.e., a single model predicts all CI agents all at once).

protests following the death of George Floyd in 2020, the Hong Kong protests by the Anti-Extradition Law Amendment Bill Movement from 2019 to 2020, and the January 6 Capitol Attack in 2021. A common pattern observed during those events is that protesters organize in a largely decentralized and leaderless fashion, while central authorities (police, governments, or authoritarian regimes) use AI (e.g., geo-location and facial recognition) to identify, punish, or suppress the protesters. This adversarial condition creates a dynamic tension between the need to coordinate at scale and the need to reduce predictability.

We adopted a well-known task in computer science, the graph coloring problem (Karp, 1972), to model coordination among the human agents. The graph coloring problem entails assigning colors to each node in a network such that no two adjacent nodes share the same color. Scheduling problems, for example, for aircraft slots, can be mapped on the graph coloring problem (Marx, 2004). Protesters, for instance, might intend to schedule participation in distributed protest in a way to avoid over- and understaffing. In our version, each agent is placed on an Erdős–Rényi random network node. We choose Erdős–Rényi random networks as they capture well the key properties one would expect of a random assembly of individuals forming spontaneous connections (Newman, Barabási, & Watts, 2006, Chapter 2). In a sequential task over multiple time steps, each agent is asked to choose one of four colors so that the selected color differs from the color selected by all its neighbors (see Fig. 1a). Both humans (Kearns, Suri, & Montfort, 2006) and algorithms (Matula, Marble, & Isaacson, 1972) can solve this problem. On the other hand, the task of the PI is to predict each color chosen by the group members correctly. The challenge for the PI thereby resembles a multiclass version of the matching pennies game (Gibbons, 1992).

All agents in the coupled system, the PI and the individual CI agents, are represented through neural networks trained through deep Q-learning (Mnih et al., 2015). Different neural network architectures have been used to investigate the decision-making of individual humans (Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021). In a similar fashion at the group level, we compare different neural network architectures as models of collective and central decision-making. We employ recurrent neural networks to model both the CI and the PI. The distinct feature of recurrent neural networks is the existence of a memory state (Yu, Si, Hu, & Zhang, 2019), which allows the agents to condition their output not only on the colors chosen in the previous timestep but also on the colors selected earlier in the sequence. Thus, both the CI and the PI can learn rich behaviors, that is, sequences of colors. We thereby allow for essentially open-ended complexity in the agent's behavior.

AI systems are typically trained centrally—even if the same models are then executed in a decentralized fashion. While the CI in our simulation only has access to the color chosen by the direct neighbors, we manipulate the ability of the PI to learn from the behavior of the CI. Fig 1.b illustrates the different conditions under consideration. A typical architecture utilized in machine learning is *centralized training and decentralized execution*. In this architecture, the AI model can access the information of the complete system during training time but only local information during execution (Su, Adams, & Beling, 2020). Such architecture corresponds to the training of a single AI, whose clones are then used to predict each human agent's behavior. We contrast such a "shared" condition with two control architectures. In the "personal" condition, each agent of the CI is paired with a personal independent agent of the PI. In the "global" condition, a single model has global information access (i.e., all past color choices by all agents) during both training and execution.

We find that through competition with a PI, the behavioral complexity of the CI increases as measured by the entropy of color sequences. The ability of the collective to coordinate is hampered but remains stable throughout the simulation. We find that competition with a single shared PI leads to a reduction in collective behavioral diversity between agents. We discuss how competition between a central PI and CI could boost the CI members' individual cultural complexity. We also discuss the danger of AI-enabled human oppression.

2. Methods

2.1. *Asymmetric cooperative-competitive game*

All 40 CI agents are placed on a network for each independent simulation, randomly generated before the simulation. We sample networks from Erdős–Rényi model with a probability for an edge creation of 0.1. For computational reasons, we reject networks with unusually highly central nodes (max degree higher than 8). Additionally, we require the network to have a single component. Agents play episodes of 20 rounds. In each round, each agent can choose one of four colors. Correspondingly, the PI has to predict the colors chosen by each of the 40 agents of the CI. For computational efficiency, all agents of the same type make their decisions simultaneously within a single round. A CI agent gets a single positive reward if its color is different from one of all neighboring agents. The PI gets a single positive reward for each color it correctly predicts. The PI receives a negative reward for each point the CI gains. Vice versa, the CI members receive a negative reward for each point the PI gains, thereby obtaining a zero-sum game.

2.2. Multi-agent reinforcement learning

We utilize deep Q-networks (Mnih et al., 2015) for all agents. Each agent is represented by a neural network. We optimize the policy using an Root Mean Squared Propagation (RMSprop) optimizer with a learning rate of 0.001. Future rewards are discounted by a factor of 0.8. At the end of each episode, we train a batch sampled from the last 500 episodes. We update the target network every 200 episodes.

For the CI, we use *individual* neural networks for each agent. These models receive the colors of the focal agent and all its neighbors as input and output of the agent's next color. We use a sandwich of a single gated recurrent unit (Chung, Gulcehre, Cho, & Bengio, 2014) enclosed by two fully connected layers.

For the PI, we contrast three different architectures and a random control. In the *personal* condition, we pair each CI agent with a unique personal PI model. In this condition, the architecture of the PI mirrors the CI. In the *shared* condition, the individual neural networks remain the same; however, the PI interacting with each CI agent shares the same model, that is, we use the same model to predict each CI agents' color independently. This setup corresponds to a central learning and decentral execution (Su et al., 2020). During training, a single central model is trained using data from the interaction with all agents. However, when making a prediction about a focal agent's next color, the PI can only use the same local information as in the personal case, namely, the previous color of the focal agent and its neighbors. In the *global* condition, a single neural network is used to predict the colors of all nodes. The network receives as input the colors of all nodes. We use in sequence a fully connected layer, a gated recurrent unit, and two fully connected layers. Across conditions, we use 100 neurons in each layer, and all linear layers use rectified linear unit (ReLu) activations. Finally, we also report the results of a random *control* using a PI making uniformly random predictions.

2.3. Behavioral metrics

We calculate the rolling average of the prevalence of CI agents correctly coordinating with their neighbors and the average number of CI agents correctly predicted by the PI. Both measures, with inverse signs, reflect the objectives of the two agent types. CI agents seek to increase coordination and decrease being predicted. The PI agents seek the inverse.

Additionally, we report measures of behavioral complexity and diversity of the CI agents. Toward that goal, we analyze the color sequences used by the individual CI agents. In particular, we analyze the distribution of sequences of four consecutive colors used by the agents. We split the sequence of 20 actions that comprise an episode into tuples of four colors, such as "RGRG" for an agent that alternates between red and green. We repeat the process for

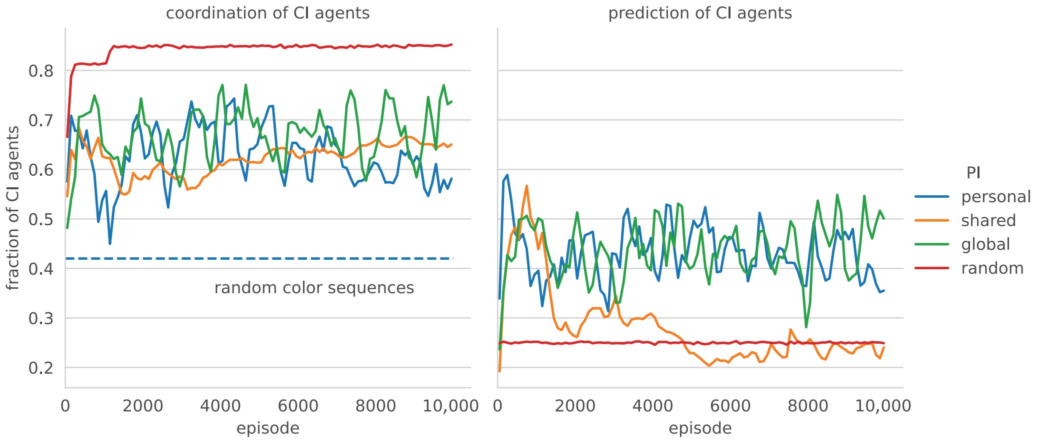

Fig. 2. On the left, we show the fraction of CI agents correctly coordinating with their neighboring nodes, that is, choosing a different color from all their neighbors. The right panel shows the fraction of agents that the PI accurately predicts. A PI with personal models for each agent is depicted in blue. A PI with all agents sharing the same model is depicted in orange. A global model predicting all agents simultaneously is shown in green. For reference, we also show random predictions in red.

the four possible offsets. We then count the frequency of unique tuples within bins of 100 episodes and calculate the corresponding entropy. A lower entropy indicates that agents utilize a small number of possible sequences. For instance, an agent that uses a single color throughout the full episode will be represented by only one tuple (e.g. "RRRR" for using red only). A higher entropy indicates greater complexity in the sequences of colors used by individual CI agents. For instance, a sequence alternating between two colors will correspond to at least two tuples (e.g. "RGRG" and "GRGR" for green and red). We use a second metric to measure the diversity of sequences used by different CI agents. Toward that goal, we calculate their Jensen–Shannon divergence, that is, the average Kullback–Leibler divergence between the sequence distribution of each CI agent and the mean sequence distributions of all agents. A larger Jensen–Shannon divergence corresponds to a larger diversity in the sequences of colors chosen by different CI agents.

3. Results

Here, in the main text, we show the results for one set of parameters (a learning rate of 0.001 and neural networks with a hidden size of 100 units). Additionally, in the Supporting Information, we report a set of different learning rates and network sizes.

3.1. Coordination and prediction

Fig. 1,2 shows the fraction of nodes that correctly coordinate with their neighbors and the fraction of nodes correctly predicted by the PI. Due to the presence of only four colors, the baseline accuracy of a random predictor is 25%. The PI is most successful in predicting the color choices in the global condition with an accuracy of 46% in the last 1000 episodes. In the personal condition, the accuracy is slightly lower at 40%. In the shared condition, the PI starts with similar accuracy, compared to the other conditions in the first 1000 episodes. However, after the first 1000 episodes, accuracy drops considerably and is eventually slightly below the chance level at 24%.

Although the CI can reduce the chance of being predicted in the shared condition, the PI is still considerably limiting the ability of CI to coordinate. The baseline for coordination of random agents is 42%. When predictions are random, CI agents lack any competitive pressure. Under these conditions, 85% of CI agents achieves correct coordination on average in the last 1000 episodes. The proportion of coordinated agents drops to 65% when facing a single cloned PI (shared condition). Coordination rates in the personal and the global conditions are similar, 58% and 69%, respectively.

Qualitatively, the simulations reveal distinct oscillations and fluctuations in both the level of coordination and the accuracy of predictions. These oscillations are less distinct in the shared weight condition than in the other conditions. A possible explanation could be that a coupling of the predictive models via shared weights might be dampening the system's oscillations.

3.2. Color sequences

Over the simulation period, individual agent color sequences appear increasingly random (see Fig. 3). At the beginning of the simulation (left and middle panels), individual agents mostly stick with a single color during one episode. In contrast, in the case of a shared PI, repetitions of the same color do not appear to be above chance toward the end of the simulation (right). In the case of the personal, global, and random PI, some agents are still using a single color throughout a full episode. However, in the personal and global conditions, agents typically choose different colors across episodes. In contrast, a number of agents in the random condition choose the same favorite color throughout the simulation.

We validate this observation by constructing tuples of four consecutive colors from the sequences of the individual color choices of the CI agents. The frequency distribution of these sequences is displayed in Supporting Information Fig. S4. Across all types of PI opponents, CI agents initially used colors repetitively. Correspondingly, tuples of four identical colors (e.g., RRRR) are overrepresented. In the personal and the shared condition, the frequency of the different sequences equalizes throughout the simulation, approaching a uniform distribution (see Fig. S4). This process is faster in the condition with a shared PI, and the CI agents' color choices appear uniformly distributed toward the end of the simulation.

CI agents escape prediction by using complex color sequences within an episode and adaptation across episodes. For instance, CI agents in the random and the global condition predominantly rely on a single color within a full episode. Yet, to reduce predictability, CI agents change their colors regularly across episodes. In Fig. S3, we show the Jensen–Shannon divergence across episodes as a measure of this adaptation. In the personal condition, agents show adaptation. Meanwhile, in the shared condition, agents display, to varying degrees, both episode-over-episode adaptation and an increase in sequence complexity.

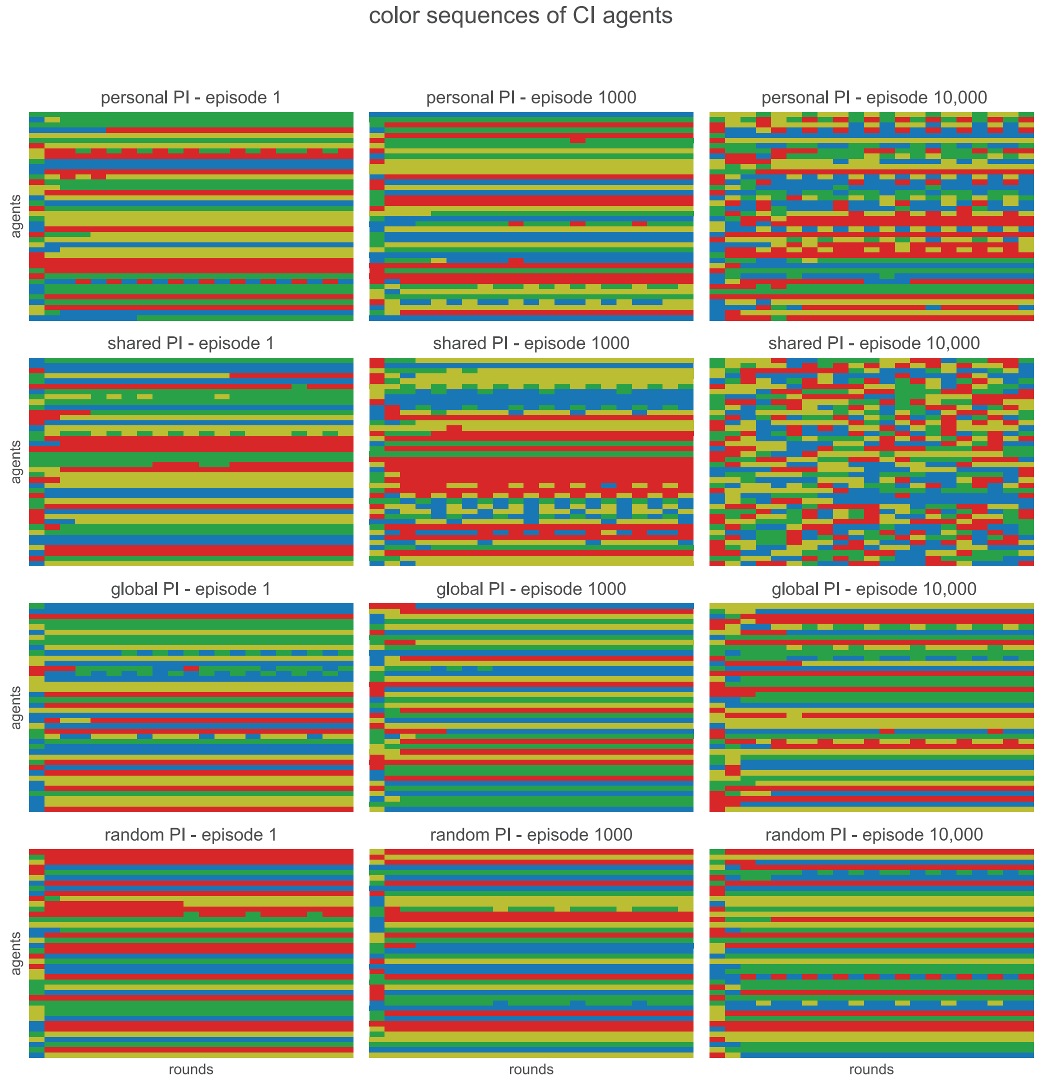

Fig. 3. Snapshots of the colors chosen by the CI agents facing a personal, shared, global, and random PI. Agents stick to individual colors at the beginning of the simulation. In particular, in the shared and, to a lesser degree, in the personal and global conditions, agents change colors frequently toward the end of the simulation. Each row corresponds to a single agent within the simulation. Columns represent the 20 rounds that comprise an episode. Note that when facing a random PI, many agents kept a single color throughout the simulation.

3.3. Increase of entropy

We then measured behavioral complexity in terms of the sequence entropy. Complexity tends to increase over time in cultural and biological evolutionary processes (Marshall et al., 2021). We calculated the average entropy of the four-color tuples across all agents as a measure of behavioral complexity. The left panel in Fig. 4 depicts the evolution of the entropy

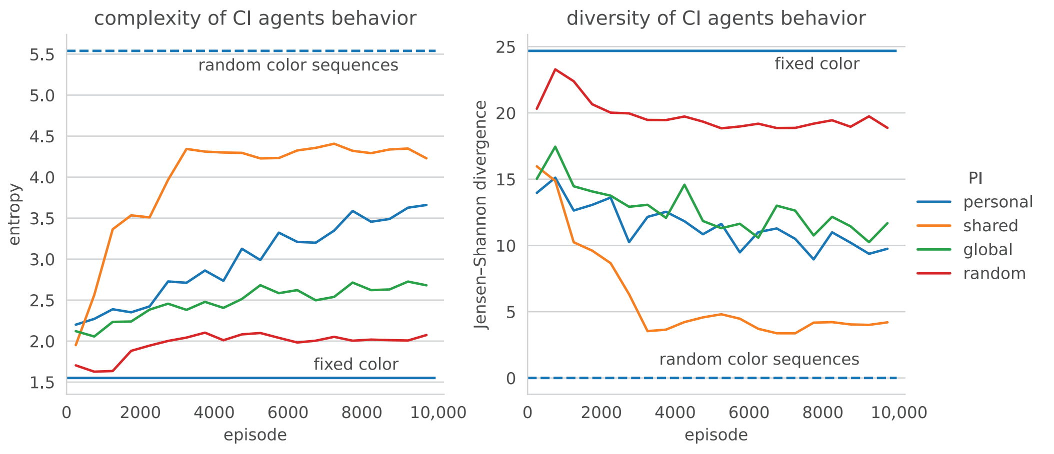

Fig. 4. On the left, we show the entropy of tuples of four sequential colors chosen by the CI agents. A higher entropy can be interpreted as a larger diversity in the chosen color sequences. On the right, we show the Jensen–Shannon divergence of the same tuples between all agents. A lower value can be interpreted as a larger similarity of the color sequences chosen by different agents. We added as reference a solid black line for agents using a fixed color across all rounds and a dashed black line for agents using a random color each round.

over the learning period. In the case of random predictions, we observe only a mild increase in complexity with an average entropy of 2.1 in the last 1000 episodes. The entropy of CI agents in the global condition is slightly higher (2.7). In the personal condition, entropy constantly increases over the training period, leading to an entropy of 3.6. We found the highest entropy in simulations with a single shared model predicting the action of each agent. In this case, the entropy increased rapidly in the first 3000 episodes and reached a value of 4.3 eventually.

3.4. *Reduction of diversity*

The entropy analysis showed that at the individual level, the competitive pressure created by the PI drives CI agents to increase their behavioral complexity. Increased behavioral complexity, in turn, reduces the performance of the PI to levels comparable to the CI adopting a random behavior (Fig. 2b). Yet, random behavior would not have benefited the group because of the evolutionary pressure on individuals created by the need to coordinate. To test the effects of increased behavioral complexity at the individual level on group-level coordination, we investigated the diversity of the color sequences used by different agents as measured by their Jensen–Shannon divergence. In the global, personal, or random conditions, the similarity of behavior between agents was relatively stable throughout the learning period. Most strikingly, the exposure to a PI with shared weights considerably reduced the divergence in behavior between agents (Fig. 4b). In the shared condition, the divergence is reduced from 15.4 in the first 1000 episodes to 4.1 in the last 1000 episodes. It is worth noting that a similar reduction in diversity is observed in the personal and global PI conditions when the learning rate is increased (see Fig. S2). These results suggest that when CI agents face low competitive pressure, they can rely on choosing a color that works given their neighbors' choices and sticking to those. This repetitive behavior also means that different agents will likely choose different colors, so diversity is large. On the other hand, when competitive pressure is high, individual agents stop relying on specific colors and need to alternate between colors to reduce predictability. Notice, however, that while entropy increases at the individual level (individuals alternate between colors more frequently), behavioral diversity at the group level reduces. Moreover, the coordination ability is on par in all conditions (with the exception of the random control). This result suggests that rather than increasing the randomness of their choices, CI agents in the shared condition evolve complex, behavioral patterns to coordinate while keeping individual actions unpredictable.

4. Discussion

In this work, we investigated an adversarial system between a predictive algorithm and a collective of decentralized learners. We adapted the graph coloring and matching pennies problems and combined them into a new zero-sum game. The predictive algorithm and the collective were both modeled as agents using neural networks and trained using reinforcement learning. We found that the ability of decentralized agents to coordinate was relatively independent of the architecture of the predictive algorithm. On the other hand, the architecture strongly impacted the complexity and diversity of emergent behaviors. In the following, we summarize our results and suggest possible interpretations.

First, competition with PI leads CI agents to innovate. Our simulation shows that competition with another adversarial learning system can lead to increasingly complex behaviors as measured by entropy. In economic and organizational ecosystems, competition drives innovation (Nelson, 1985). In the last decade, people have developed AI systems that have superhuman capabilities in an increasing number of fields. Thus, in a hybrid ecosystem comprising both humans and AI agents, competition might drive mutual innovation. One example of this phenomenon is in the game of Go. AlphaGO was the first AI to beat professional human players in the game (Silver et al., 2016). Subsequently, professional human players improved their gameplay by playing against an open-source version of the AlphaGo algorithm (Choi, Kim, Kim, & Kang, 2021).

Second, PI drives CI agents to use increasingly complex behavioral patterns to resolve the dilemma of coordination and predictability. A trivial solution to reduce predictability is for agents to choose colors at random. However, such random behavior does not allow for coordination between agents. We find in our simulations that agents learn increasingly complex behavioral patterns. In scenarios where agents compete against a shared PI, the entropy of their behavior approaches that of random actions, while still enabling coordination. In such situations, using complex sequences can be an effective behavioral pattern to reduce predictability and gain an advantage over the shared PI. On the other hand, when competing against a global PI, a simple adaptation of a behavior, such as changing the dominant color, during training is often sufficient. The difference in behavioral patterns might arise from the global PI's ability to personalize its predictions, allowing it to exploit patterns overrepresented among individual CI agents. In contrast, the shared PI lacks this capability and is more likely to focus on identifying overrepresented sequence patterns, such as repeated use of the same color. Consequently, CI agents may develop increasingly complex patterns. We measured complexity in our study by the entropy of tuples representing sequentially used colors by each agent. Future studies should explore the use of "assembly theory" as a tool to measure behavioral complexity. This approach—originally proposed to study molecular complexity (Marshall et al., 2021)—studies increasing complexity by modeling the compositionality of units and subunits via basic operations.

Third, competition between diverse CI agents and a central PI aligns the CI agents' behavior. This result is not trivial. It is unclear why increasing behavioral alignment (reduced diversity) emerges from the competition of distributed learners against a centralized learner. One might assume that by diversifying individual behaviors, the predictability by a centralized agent could be reduced. In our simulations, we found the opposite. When we paired CI agents in our simulations with clones of a single PI, the diversity of their behavior decreased. We used a derivative measure of Kullback-Leibler divergence to measure diversity. Diversity decreases suggest agents learned to match their distributions of color frequencies. This result, combined with the observed increased entropy, indicates that agents in the condition with a shared PI evolved a shared behavioral code (the color frequency distribution) to achieve both coordination with their peers and behavioral unpredictability.

If confirmed, these findings could have interesting implications for computational models and civil society. The ability to instantly share learnings across the entire system could be a potential advantage of AI. We found that decentralized agents were better able to reduce predictability when the predictive algorithm shared its weights across all agents. The shared PI in our study represents the typical AI architecture of centralized training and decentralized execution used in everyday life. In fact, academia and industry are relying on a relatively small number of so-called foundation models (e.g., BERT, GPT-3, AlexNet; Bommasani et al., 2021). Such foundation models are centrally pre-trained but locally executed by individual applications and fine-tuned based on individual users. We found that the decentralized learner struggled to escape the prediction when the predictive algorithm was able to adapt to individual agents in the global and personal condition. This finding suggests that hyper-personalization of algorithms could have an increased danger of oppression, compared to the utilization of general-purpose algorithms.

The popular *Terminator* movie series mainstreamed the trope of an epic war between the human race and a superhuman AI. While modern AI is not as powerful as Skynet, it is regularly used by central authorities (e.g., police, governments, and authoritarian regimes) to identify protesters and suppress civil uprisings, thanks to geo-location, face recognition, and AI profiling. As highlighted in the previous paragraph, AI applications rely on centrally trained models. While in the *Terminator* movies, human resistance centers around its leader (John Connor), real-world protests in recent history mainly relied on large-scale coordination of distributed leaderless movements via Twitter and other social media. Recent notable examples include Black Lives Matter's protests following the death of George Floyd in 2020, protests in Hong Kong following the Anti-Extradition Law Amendment Bill from 2019 to 2020, and the January 6 Capitol Attack in 2021. Our study focuses on the tension between centralized AI and decentralized distributed coordination, using established tools from computer science, such as multi-agent reinforcement learning, as well as familiar games like coloring graphs and matching pennies.

An additional contribution of this work is introducing a novel system to study cultural evolution via competitive dynamics between centralized and distributed learning systems. Cultural evolution is a dynamic process enabled by social learning and competitive pressures (Boyd & Richerson, 1988). Coordination (e.g., emergence of norms) and behavioral complexity are hallmarks of culture (Muthukrishna, Shulman, Vasilescu, & Henrich, 2014). Cumulative culture can emerge in a largely decentralized fashion from local interactions among social learners (Sasaki & Biro, 2017). Unlike other human cultural artifacts, AI learns from experience without further human intervention. Our findings indicate that when a technological learning system exists in a competitive relationship with distributed learners, it can speed up behavioral, and thus cultural, evolution by exerting competitive pressure on human learners. In our study, the emergence of behavioral complexity was driven by the need for less predictable behavioral patterns to achieve coordination.

While competition in our simulation led to increased behavioral complexity, this should not be taken as evidence that an antagonistic relationship between humans and AI is desirable. Scholars, policymakers, and regulators, in particular the European Commission, are working to define AI's legitimate scope, with some suggesting a legal and technical right to opt out of such exposure (Dwork, 2006; Zuboff, 2019). More work is needed to understand the short- and long-term consequences of predictive algorithms in the real world, particularly the risks associated with further marginalizing already marginalized groups. The algorithmic fairness of often-centrally trained proprietary general-purpose algorithms is an actively welcomed field of research (Ali et al., 2019; Angwin, Larson, & Mattu, 2016; Bommasani et al., 2021; Buolamwini & Gebru, 2018; O'Neil, 2016). In conclusion, our findings show an increase in cultural complexity in a very limited testing environment. They do not provide evidence for negative or positive real-world implications of human–AI competition. As a result, they should be interpreted with caution.

We would like to stress that the findings presented in this work might depend on the particular hyperparameters selected. Reinforcement learning using deep Q-learning has many free parameters, such as the learning rate, the size and architecture of the neural networks, and the update frequency of the target network. Due to the dynamic and potentially chaotic nature of multi-agent systems, results can vary dramatically based on the chosen parameters. Some experimental settings, such as the architectures used, were made based on best judgment. We present additional results in the Supplementary Material with different learning rates and model complexities. For instance, when using a faster learning rate, differences in predictability between the different conditions vanish (see Fig. S2), and sequence complexity in the individual condition matches the shared condition (see Fig. S3). This suggests that our results might be limited to systems of intermediate adaptability.

We encourage further research into competitive games that feature complex network topologies. This work is based on matching pennies and graph coloring as textbook examples for mixed strategies and graph coordination, respectively. Yet, other games, such as Stackelberg Security Game, specifically describe the asymmetry between an attacker and a defender (Sinha, Fang, An, Kiekintveld, & Tambe, 2018) that could arise between protesters and a security force. In the current study, we explore the role of information asymmetry in a rudimentary manner. While we modulate on the AI side between global and local information, allowing the CI agents to communicate by encrypted channels or partially revealing their plans could allow for new strategies to reduce predictability (Xu, Rabinovich, Dughmi, & Tambe, 2015). In this work, we present the dynamics on an Erdős–Rényi network with 40 nodes. In the real world, protests might involve hundreds of thousands of individuals. Our method does not scale to such network sizes, and a combination with more traditional agent-based modeling might be required to investigate the dynamics at such scales.

Furthermore, the approach of modeling human behavior purely as utilitarian has its known limitations. In our model, agents learn to maximize discounted expected reward through temporal difference. While there is experimental evidence of the importance of temporal difference, as used in Q-learning, in the brain (Schultz et al., 1997), human behavior is repeatedly found to be non-utilitarian as described by prospect theory and loss aversion (Kahneman 2011). Future studies might investigate the role of non-utilitarian behavior on the collective dynamics presented in this work.

This study aimed to examine the dynamics between centralized and decentralized learners when coupled in an adversarial system. Future studies will need to replicate the findings presented here and measure the effects of the game, the network size, the model and training parameters on these behaviors.

In conclusion, our findings indicate that when a centralized learning system competes with a distributed population of learners, the resulting dynamics can be surprisingly complex, leading to increased behavioral complexity and reduced behavioral diversity.

Acknowledgments

Manuel Cebrian was partially supported by the Ministry of Universities of the Government of Spain, under the program “Convocatoria de Ayudas para la recualificación del sistema universitario español para 2021–2023, de la Universidad Carlos III de Madrid, de 1 de Julio de 2021.”

Open access funding enabled and organized by Projekt DEAL.

Open Research Badges

This article has earned Open Materials badge. Materials is available at https://github.com/LBrinkmann/ai_ci.

References

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *3* (CSCW), 199.

Angwin, J., Larson, J., & Mattu, S. (2016). Machine bias. Ethics of data and analytics. https://doi.org/10.1201/9781003278290-37

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula. arXiv. http://arxiv.org/abs/1909.07528

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv. https://arxiv.org/abs/2108.07258

Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process* (reprint ed.). Chicago, IL: University of Chicago Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st conference on fairness, accountability and transparency* (Vol. 81, pp. 77–91). New York: PMLR.

Chen, L., Ning, H., Nugent, C. D., & Yu, Z. (2020). Hybrid human-artificial intelligence. *Computer*, *53*(8), 14–17.

Choi, S., Kim, N., Kim, J., & Kang, H. (2021). How does AI improve human decision-making? Evidence from the AI-powered Go program. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3893835

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2015). Gated feedback recurrent neural networks. In International conference on machine learning (pp. 2067-2075). PMLR.

Coupé, C., Oh, Y. M., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, *5*(9), eaaw2594.

Couzin, I. (2007). Collective minds. *Nature*, *445*, 715. https://doi.org/10.1038/445715a

Dawkins, R., & Krebs, J. R. (1979). Arms races between and within species. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character. Royal Society*, *205*(1161), 489–511.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643.

Dwork, C. (2006). Differential privacy. In International colloquium on automata, languages, and programming (pp. 1-12). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/11787006_1

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.

Gibbons, R. S. (1992). *Game theory for applied economists*. Princeton, NJ: Princeton University Press.

Gillings, M. R., Hilbert, M., & Kemp, D. J. (2016). Information in the biosphere: Biological and digital worlds. *Trends in Ecology & Evolution*, *31*(3), 180–189.

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, *24*(11), 873–883. https://doi.org/10.1016/j.tics.2020.09.001

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology*, *12*(3), 123–135.

Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Cambridge, MA: Harvard University Press.

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, *37*(1), 81–88.

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed., Vol. 55). New York: Farrar, Straus and Giroux.

Karp, R. M. (1972). Reducibility among combinatorial problems. In R. E. Miller, J. W. Thatcher, & J. D. Bohlinger (Eds.), *Complexity of computer computations: Proceedings of a symposium on the complexity of computer computations, held March 20–22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New*

York, and sponsored by the Office of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department (pp. 85–103). Boston, MA: Springer.

Kearns, M., Suri, S., & Montfort, N. (2006). An experimental study of the coloring problem on human subject networks. *Science*, *313*(5788), 824–827.

Kittur, A., Lee, B., & Kraut, R. E. (2009). Coordination in collective intelligence: The role of team structure and task interdependence. In D. R. Olsen & Association for Computing Machinery (Eds.), *Proceedings of the SIGCHi conference on human factors in computing systems* (pp. 1495–1504). New York: ACM.

Konečný, J., Brendan McMahan, H., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. arXiv. https://arxiv.org/abs/1610.05492

Köster, R., Hadfield-Menell, D., Everett, R., Weidinger, L., Hadfield, G. K., & Leibo, J. Z. (2022). Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(3), e2106028118. https://doi.org/10.1073/pnas.2106028118

Leibo, J. Z., Hughes, E., Lanctot, M., & Graepel, T. (2019). Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. arXiv. https://arxiv.org/abs/1903.00742

Liu, S., Lever, G., Merel, J., Tunyasuvunakool, S., Heess, N., & Graepel, T. (2019). Emergent coordination through competition. arXiv. https://arxiv.org/abs/1902.07151

Malone, T. W. (2018). *Superminds: The surprising power of people and computers thinking together*. New York: Little, Brown Spark.

Marshall, S. M., Mathis, C., Carrick, E., Keenan, G., Cooper, G. J. T., Graham, H., Craven, M., Gromski, P. S., Moore, D. G., Walker, S. I., & Cronin, L. (2021). Identifying molecules as biosignatures with assembly theory and mass spectrometry. *Nature Communications*, *12*(1), 3033.

Marx, D. (2004). Graph colouring problems and their applications in scheduling. *Periodica Polytechnica, Electrical Engineering*, *48*(1-2), 11–16.

Matula, D. W., Marble, G., & Isaacson, J. D. (1972). GRAPH COLORING ALGORITHMS††This research was supported in part by the Advanced Research Projects Agency of the Department of Defense under contract SD-302 and by the National Science Foundation under contract GJ-446. In R. C. Read (Ed.), *Graph theory and computing* (pp. 109–122). New York: Academic Press.

Mesoudi, A. (2017). Pursuing Darwin's curious parallel: Prospects for a science of cultural evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(30), 7853–7860.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Muthukrishna, M., Shulman, B. W., Vasilescu, V., & Henrich, J. (2014). Sociality influences cultural complexity. *Proceedings. Biological Sciences /The Royal Society*, *281*(1774), 20132511.

Nelson, R. R. (1985). *An evolutionary theory of economic change*. Cambridge, MA: Harvard University Press.

Newman, M. E., Barabási, A. L. E., & Watts, D. J. (2006). *The structure and dynamics of networks*. Princeton, NJ: Princeton University Press.

Niazi, M., & Hussain, A. (2011). Agent-based computing from multi-agent systems to agent-based models: A visual survey. *Scientometrics*, *89*(2), 479–499.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (Vol. 24). New York: Crown.

Pescetelli, N. (2021). A brief taxonomy of hybrid intelligence. *Forecasting*, *3*(3), 633–643.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214.

Reia, S. M., Amado, A. C., & Fontanari, J. F. (2019). Agent-based models of collective intelligence. *Physics of Life Reviews*, *31*, 320–331.

Riedl, C., Kim, Y. J., Gupta, P., Malone, T. W., & Woolley, A. W. (2021). Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(21), e2005737118. https://doi.org/10.1073/pnas.2005737118

Russell, S. J., Norvig, P. (2010). *Artificial intelligence a modern approach*. London.

Sandholm, T. W., & Crites, R. H. (1996). Multiagent reinforcement learning in the Iterated Prisoner's Dilemma. *Biosystems*, *37*(1-2), 147–166.

Sasaki, T., & Biro, D. (2017). Cumulative culture can emerge from collective intelligence in animal groups. *Nature Communications*, *8*, 15049.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Sinha, A., Fang, F., An, B., Kiekintveld, C., & Tambe, M. (2018). Stackelberg security games: Looking beyond a decade of success. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, Stockholm, Sweden.

Su, J., Adams, S., & Beling, P. (2021). Value-decomposition multi-agent actor-critics. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 13, pp. 11352-11360)

Surowiecki, J. (2005). *The wisdom of crowds*. New York: Anchor.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). Cambridge, MA: MIT Press.

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

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.

Xu, H., Rabinovich, Z., Dughmi, S., & Tambe, M. (2015). Exploring information asymmetry in two-stage security games. *Proceedings of the AAAI Conference on Artificial Intelligence*, *29*(1). https://doi.org/10.1609/aaai.v29i1.9290

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, *31*(7), 1235–1270.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power: Barack Obama's books of 2019*. London: Profile Books.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab of this article:

Supplementary Material