Beyond the Matrix: Experimental Approaches to Studying Cognitive Agents in Social-Ecological Systems

Uri Hertz[1], Raphael Koster[2], Marco A. Janssen[3], and Joel Z. Leibo[2]

[1]Department of Cognitive Sciences, University of Haifa, Haifa, Israel
[2]Google Deepmind, London, UK
[3]School of Sustainability, Arizona State University, Tempe, USA

December 7, 2024

Accepted version, to be published in Cognition.

1 Abstract

Social-ecological systems, in which agents interact with each other and their environment are important both for sustainability applications and for understanding how human cognition functions in context. In such systems, the environment shapes the agents' experience and actions, and in turn collective action of agents changes social and physical aspects of the environment. Here we review current investigation approaches, which rely on a lean design, with discrete actions and outcomes and little scope for varying environmental parameters and cognitive demands. We then introduce multi-agent reinforcement learning (MARL) approach, which builds on modern artificial intelligence techniques, which provides new avenues to model complex social worlds, while preserving more of their characteristics, and allowing them to capture a variety of social phenomena. These techniques can be fed back to the laboratory where they make it easier to design experiments in complex social situations without compromising their tractability for computational modeling. We showcase the potential MARL by discussing several recent studies that have used it, detailing the way environmental settings and cognitive constraints can lead to the emergence of complex cooperation strategies. This novel approach can help researchers bring together insights from human cognition, sustainability, and AI, to tackle real world problems of social-ecological systems.

2 Introduction

Some of the biggest challenges facing humanity, such as sustainability, pandemics, and conflict resolution, are global in nature but require coordinated effort of many local communities to address them. This requires understanding of the complex inter-dependent relations between humans and their social and biophysical environment as a social-ecological system (Ostrom, 2009; Schill et al., 2019). In social-ecological systems, the agents' experience and actions are shaped by the affordances provided by their physical environment, such as spatial and temporal constraints, resources and dangers, and by affordances created by their social setting including those of conventions, norms, reciprocity, and reputation. The environment impacts the decision setting faced by individuals and interacts with their cognitive processes, such as learning, memory, planning and attention. Social-ecological systems may thus involve complex couplings between behavioral, social, physical, and ecological variables since the environment where agents act is itself changed by their actions.

Social-ecological systems are not just complex because of the number of agents or the size the system encompasses. Their complexity arises from the variety of behavioral affordances the environment provides, and the different ways each agent can combine them in space and time to form a behavioral pattern, and the non-linear feedback loops and interactions between these patterns and the social and physical environment (Levin et al., 2013). Empirical research on shared resource governance has demonstrated the importance of the spatial and temporal context of the resource. They impact the way groups may be able to cooperate and coordinate on resource governance (Folke et al., 2007; Young, 2002). Studies of long-lasting social-ecological systems, such as fisheries, show that institutional rules are mainly based on where, when, and how to harvest, not how much to harvest (Schlager et al., 1994; J. A. Wilson et al., 1994). This suggests that an understanding of the fit between institutions and ecology needs to understand how human activities can match the temporal and spatial dynamics of specific resources. It is also important to take into account the cognitive processes and motivation structures of the individuals that operate in, and interact with, their environment (Donges et al., 2020; Schill et al., 2019). Thus individual cognitive abilities like memory and attention shape collective behavior. For instance, they may determine which part of the environment is perceived by individuals, and therefore which environmental signals are likely to affect behavior. Norms and institutions governing how individuals use environmental resources may be seen as evolving through a kind of group-level trial and error process (Ostrom, 1998). Thus an account of how individuals learn, and how the learning of individuals constitutes group-level learning, is critical in the study of social-ecological systems.

The study of social-ecological systems uses many qualitative and quantitative approaches as demonstrated in various methodological textbooks in this field (Biggs et al., 2021; Poteete et al., 2010). Researchers use questionnaires and field experiments to understand the dynamics of social norms (Blair et al., 2019), field work and geographic data to study cultural and environmen-

tal differences (Bansak et al., 2018). Computational models use multi-agent simulations (J. Wilson et al., 2007) and formal differential equations (Elsawah et al., 2017) to predict and uncover factors shaping social-ecological dynamics. These approaches have their individual merits and limitations, such as tracking real-world settings and problems, or allowing formal solutions to predict social dynamics. However, in this review we narrow in on two approaches to focus on collective action in social dilemmas appearing in social-ecological systems. The first approach is the use of simple matrix games to do controlled experiments for relevant social dilemmas such as common pool resources and public goods. We contrast this with an emerging approach called MARL (Multi-Agent Reinforcement Learning) which disaggregates the decisions from a matrix game into many sub decisions in a dynamic environment using computational learning agents. We review some of the history related to the emergence and use of matrix games in behavioral science, and try to delineate their limitations in explaining and studying social-ecological systems, i.e. which dimensions they can easily capture and which they can not. We then discuss in detail how to experimentally study complex social-ecological systems using a particular method, MARL, which works synergistically with laboratory-based experimental designs involving environment simulations that have complex spatial and temporal dynamics. We review results from taking this approach, and conclude by discussing open problems where these methods could be applied in the future.

2.1 Experimental Social Psychology Approach

A long tradition in experimental social psychology addresses the complexity of human behaviour and its interaction with environmental settings. In the early literature, the aim was to design experiments to be as similar as possible to real-life scenarios. It was common to employ elaborate experimental designs involving theatre-like productions with actors, sets, and scripts. For instance, to study conformity, Asch employed multiple confederates to give an obviously wrong answer in a simple perceptual task to study one participant Asch (1951), and Milgram studied deference to authority using electrical equipment and confederates playing an electrocuted victim Milgram (1963). In Zimbardo's Stanford prison experiment, an entire group of participants were put in an isolated perimeter designed to simulate a prison, and given roles of prisoners and guards, to examine the stability of personality traits (Zimbardo, 1969). These experiments were amongst the most influential in psychological science, and left a memorable impression on both the scientific community and the general public.

While this elaborate approach had great impact, and led to many scientific insights, it also has a number of limitations. First, relying on such elaborate design makes such experiments very hard to replicate and very sensitive to experimenter effects (Camerer et al., 2018; Nosek et al., 2022). Confederates may influence participants' behaviour in subtle ways, and these influences may accumulate when designs rely on multiple confederates and circumstantial effects. Another problem is scalability. In many cases a lot of effort was invested to produce one data point,which limits the number of participants one can feasibly test

in each study, and restricts the feasibility of carrying out such experiments with varied populations across the world or in non-university settings like hospitals. In addition, many studies used bespoke outcome measurements, making it very hard to pool results across studies and evaluate effects using meta-analysis, thus making it hard to build larger multi-lab research programs. This is especially problematic when effect sizes are small since it makes false positive results more likely, and thus greatly restricts the impact of this line of work.

2.2 The matrix games approach

Behavioural game theory offers a different experimental approach to studying human social behaviour (Camerer, 2011) (Figure 1B). In a common task, the Prisoners' Dilemma, two participants choose between two strategies—to cooperate or to defect. The outcome for each participant is determined by the joint action of both, for example if both choose to cooperate then the reward to both is high, but if one defects and one cooperates, the reward obtained by the defector is higher than for the cooperator. The reward per participant per combination of actions is usually displayed in a matrix, hence the name 'matrix game'. Game theoretically "simple" models are not always two-player two-action simultaneous move games like the Prisoners Dilemma. For instance they could have more players or actions may be taken sequentially. However, to simplify terminology, we include all such games in the category we call matrix games to distinguish it from the more complex games we discuss below which cannot be compactly described in a formalism where payoffs are directly determined by joint action without additional dependence on the environment or other factors.

Matrix games became popular in the experimental study of cooperation in behavioral economics, neuroeconomics and social cognitive psychology due to a number of appealing characteristics. First, they offer a concise way of capturing the idea of a social dilemma—a tension between individual and group rationality. Instead of theatrically elaborated schemes with confederates, they focus on quantifiable incentives, which they view as determined by the "rules of the game", and regard as the sole determinant of rational strategy. Second, matrix games are flexible, as simply changing the outcomes associated with actions can produce different kinds of dilemmas, allowing the study of cooperation, coordination and asymmetric relationships. Third, because the problem is represented in a matrix, which is understood the same way by different labs, it allows consistent deployment in different settings and in different populations. For example, this facilitates testing participants around the world, and in hard-to-reach places, with the same experimental design (Henrich et al., 2001).

One of the most compelling features of the matrix game approach is that game theory provides formal solutions which define the optimal strategy for participants. For example, for a given payoff matrix and set of beliefs about co-player intentions, it is possible to determine optimal strategies which can then be used as an idealized reference to which human participants' behaviour can be compared. This is a powerful feature as it allows experimenters to draw predictions based on theory (Hoffman & Yoeli, 2022). When theoretical predictions

fail to capture observed behaviour, it is possible to refine their assumptions, for example by introducing subjective utility functions. For example, Fehr and Schmidt used a social utility concept that assumes people don't like unequal payoffs, i.e. inequity aversion, and showed it changed what is optimal to do in a prisoners' dilemma Fehr & Schmidt (1999b). Such interaction between theoretically derived predictions and behavioural experiments allows researchers to better formalize and reproduce theoretical insights and predictions.

While matrix games have been a very successful tool for studying human social behaviour, their lean structure constrains their ability to address complex features of socio-ecological systems, such as cognitive, spatial and temporal dimensions of social behaviour. Actions in matrix games, such as cooperation and defection, are manifested in discrete choices between two (or more) options, and usually directly related to monetary allocations (Figure 1B). This leaves matrix games unable to capture the way cooperation manifests in cases where actions have a substructure like spatial and temporal complexity. For example, cooperation may be the outcome of a sequence of actions and activities, with no discrete moment in the sequence that distinguishes cooperation from defection. In real life settings, individuals make small elementary choices like walk forward, say "hello", open the door, etc. Only taken as a whole does the entire sequence of such actions then (sometimes) constitute cooperation or defection (Janssen et al., 2010; Leibo et al., 2017).

Matrix games abstract away many of the mechanisms people rely on to achieve social coordination where the small actions are often critical. Game theory is not a neutral modeling language. It highlights back-and-forth strategic thinking, where players take into account other players' knowledge of the world and of one another's intentions to make a small number of impactful choices. However, when actions are embedded in specific physical context, people have a variety of other ways of solving problems that may not involve as much back-and-forth thinking. For example, they may use spatial cues to clarify their understanding one another's intentions without needing to think much about their mental states Freundlieb et al. (2016); Sebanz et al. (2006). Joint attention can direct joint action in cases where people occupy a shared environment, making it easier for two players to observe the actions available to others, or observe a common landmark as a basis for coordination. This means that the geography in which action takes place, and the shared attention opportunities it allows (e.g. whether agents are in the same space or not) may greatly affect their choice of behaviour, and may do so in a way that is difficult to capture with matrix games.

L. J. Savage illustrated the problem with using simple matrix games to model complex phenomena using a pair of proverbs which he presented as distinguishing 'small worlds' from 'large worlds'. A small world is one where you can always "look before you leap" and individuals may thus plan effectively. Whereas a large world is one where you must sometimes "cross that bridge when you come to it" so exploration is necessary and planning unlikely to work (Savage, 1954) (although this distinction is not strictly binary in real life). When you model a social-ecological system with a simple matrix game you assume it is

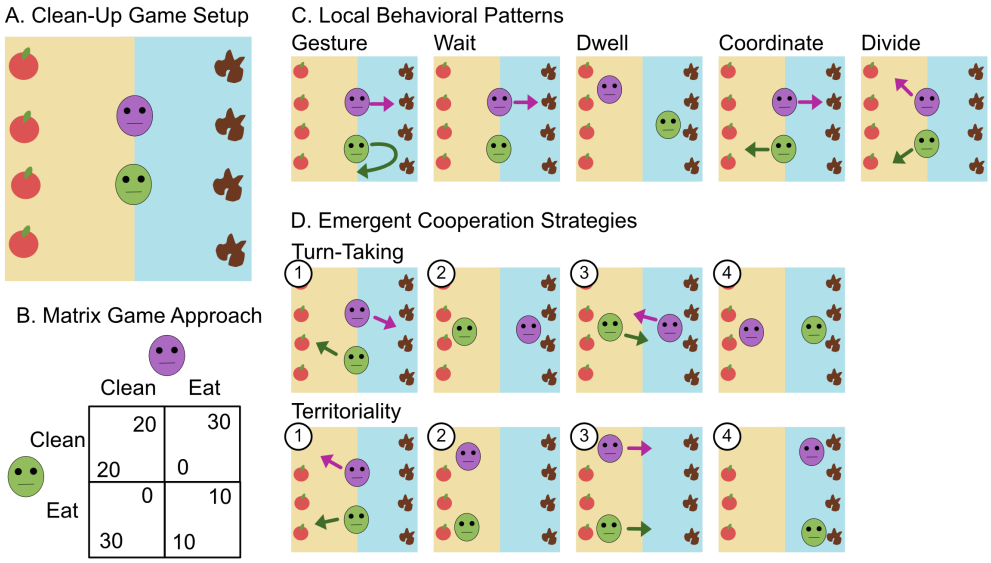

Figure 1: Approaches to studying socio-ecological systems. A. Clean Up: Apples regrow at a rate inversely proportional to accumulating pollution in the river. The pollution can be cleaned with a localized cleaning action. This environment closely mirrors the Tragedy of the Commons, or any situation with both fear of being exploited and greed motivations. The action space of the agents includes motion in any directions, cleaning pollution, and eating apples. The game can include more than two agents. Example of the game is provided at Leibo et al. (2021) (https://github.com/deepmind/meltingpot). B. The essence of the clean-up game can be captured in matrix-game, where two agents need to choose between two discrete actions: cooperation (Clean) and Defection (Eat), and their outcome is dependent on both agents' decisions. C. In MARL approach agents can learn to display many different local behavioral patterns, composed of their basic actions, which allow joint action to develop over multiple steps, such as gesturing, waiting, dwelling in one place, synchronization and mirroring of the other agent. D. The local behavioral building blocks can accumulate to form higher level strategies. For example, cooperative strategy can be to take turn in cleaning and eating, which entails inter-dependence, trust and reciprocity, or territoriality, where each agent cleans his own patch of the river and eat the apples in his territory, entailing property rights and boundaries. Local behavioral patterns and elaborate cooperative strategies are likely dependent on the environment in which agents operate, and their cognitive model, in a way that is not easy to capture in matrix games.

small in this sense, i.e. that the individuals involved may draw up forward-looking plans with some hope they could be implemented effectively. If the social-ecological system in question contains possibilities, events, affordances, resolutions, or failure modes that the individuals involved could not reasonably foresee before they occur, then it is better to adopt a model more suited to capturing social exploration in large worlds.

2.3 The multi-agent reinforcement learning (MARL) approach

In recent years a new framework to capture social-ecological complexity has emerged, building on the simultaneous development of multi-agent reinforcement learning (MARL) algorithms and their application for *in silico* simulation of human behavior in increasingly complex virtual worlds designed to incorporate principles from the social and biological sciences. The first examples in this vein to appear were mostly concerned with spatially and temporally extended analogs to the classic social dilemmas of non-cooperative game theory (Kleiman-Weiner et al., 2016; Leibo et al., 2017; Lerer & Peysakhovich, 2017; Lowe et al., 2017; Tampuu et al., 2017). However, since this approach is not naturally limited to two player games, or games with few actions, they rapidly moved beyond this origin to incorporate a wide range of social and environmental phenomena of interest (Du et al., 2023; Leibo, Hughes, et al., 2019; Nisioti & Moulin-Frier, 2020).

In this framework, models of social-economical systems consist of two interacting parts: (a) an environment simulation, and (b) its inhabitants, which can be either human players or artificial agents (Figure 2).

The environment simulation is specified by the researchers. It is an interactive program that takes inputs and returns outputs on every time-step. Formalized in reinforcement-learning terms, the inputs are the current environment state and the actions of the agents populating the environment. The outputs are the next environment state, the observations of this environment state for each agent and their respective rewards from the time-step. The environment's state influences the agents' rewards and is itself influenced by them. Examples of environment simulations include 2D worlds with various different terrains, rewarding tokens, and harming tokens. The environment simulation may also include temporal contingencies, such as the rate in which rewards are produced or terrain shifts (e.g. pollution accumulation in the riverbed, see the Clean Up environment (Fig. 4))[1]. Intricate dependencies can be encoded, for example linking reward rate with the proportion of pollution in the river. Finally, the environment also determines the specific actions available to agents, such as movement, cleaning pollution or harvesting apples. These actions, or behavioral affordances, are the building blocks of more elaborate strategies which are not

[1]For videos of the environments see; *Allelopathic Harvest*: `https://youtu.be/BbOduMGOYF4`, *Clean Up*: `https://youtu.be/TqiJYxOwdxw`, *Commons Harvest*: `https://youtu.be/lZ-qpPP4BNE`

specified by the environment. Thus, researchers can design and tweak the environment parameters to closely match the properties of the real social-ecological systems they seek to model. This is useful for testing how different interventions may affect emergent social behaviour in specific settings.

The simulator is populated by players constituted by humans or artificial cognitive models. Here, we focus on deep reinforcement-learning agents as the cognitive models of individual decision makers. Agents receive observations from their environment and output actions according to their behavioral policy. They observe the outcome of their actions, and can learn and adapt their behavior on its basis. Each agent's behavioral policy is stored in a separate neural network which is tuned gradually to increase the reward it may be expected to achieve.[2] These agents typically learn 'from scratch', i.e. the mapping from their inputs (usually pixels) to outputs (usually basic actions like, 'move forward', 'turn left') is initially random but gets refined with experience. Note that it is possible to expand this approach to capture evolutionary processes, by allowing the agents to inherit some parameters from previous rounds beyond their model weights (Jaderberg et al., 2019; Léger et al., 2023; J. X. Wang et al., 2019). As multiple agents inhabit the same simulation, their learning is interdependent; they all affect each other's observations and rewards. Since the simulated environment can be populated by human players in a very similar manner this approach makes it possible to draw behavioural predictions from the behaviour of cognitive-models trained in the environment. In addition, insights gained from humans can be used to refine the cognitive models, for example by changing their reward function, attention or working memory (Agrawal et al., 2020; McKee et al., 2021). While many deep reinforcement learning agents are simplistic in the way that they are limited to model-free reinforcement learning (Botvinick et al., 2020), the choice to use model-free techniques yields a core methodological benefit for the approach: its synthesis of visual processing in service of reward-guided action—flexible learning from scratch—offers a new arena for cognitive modeling where the experiment designer need not imbue the agent with a prior understanding of any task dynamics. Many MARL studies use the same basic architecture across myriad different environments and tasks, and need not assume prior knowledge of task dynamics (Agapiou et al., 2022; Leibo et al., 2018)[3].

The granularity of the simulation and the players makes their analysis intractable for standard game theory[4]. The state-space of the environment, observed by a stream of raw pixels, is enormous. Similarly, the actions that form the interface between the agent and the simulation are closer to low-level

[2]Note that there are different ways to implement MARL, and not all implementations have separate neural networks for each agent, some approaches share some or all neural network weights between agents (Du et al., 2023), a change which can alter the interpretation. For simplicity, we only describe the case with fully independent agents here. This is called the decentralized MARL setting.

[3]Of course model-based reinforcement learning algorithms can also be used in MARL (e.g. Silver et al. (2018))

[4]Our claim is not about game theory in general but focused on the classical "two-player two-action game toolkit.

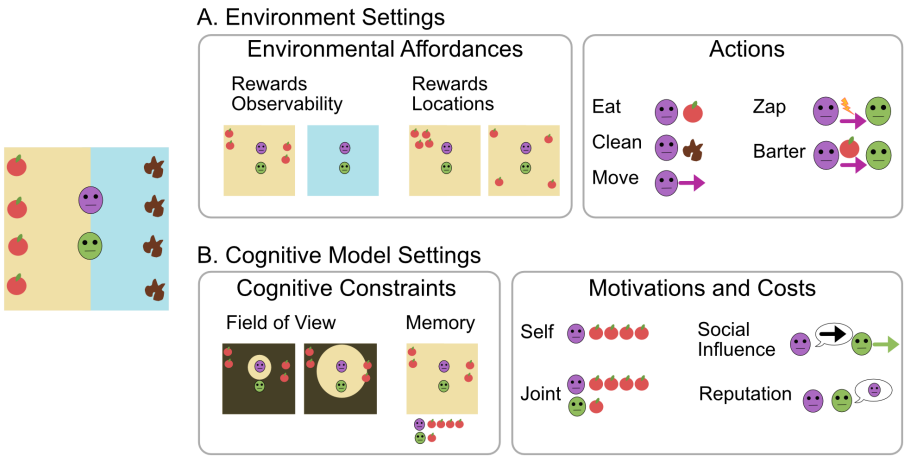

Figure 2: Two components of the MARL approach to socio-ecological systems: the environment model and the cognitive model. In our approach the experimenter can control both the environment in which the simulations or experiments take place, and the cognitive model of the agents occupying it. A. Environmental settings include the physical settings of the experiment, and the affordances it provides, such as the observability of rewards, e.g. fish vs. fruit, and their distribution in space and time. The experimenter decides about the basic set of actions available to agents, such as movement, collecting rewards and so on. By expanding the agent's action repertoire can greatly affect the local behaviors and emergent strategies learned by agents. B. The experimenter controls the cognitive model of the agents. The experimenter can control agent's cognitive constraints, for example by controlling the information presented to the agents. The experimenter can also set up different motivations by controlling the rewards available for agents—making different outcomes explicitly rewarding. When using artificial agents these settings are aimed at capturing different aspects of human cognition and mimic the way human players operate in such environments. When using human participants, these settings serve as experimental manipulations, and the motivations and cognitive constraints are inferred from behavior.

motor primitives, e.g., move forward, turn left, than they are to discrete and strategically-impactful 'cooperate' versus 'defect' actions. In this setting, high level strategic operations like cooperation must be implemented by stringing together a sequence of low-level actions suitable for the current environment state (Leibo et al., 2017). These two features of the MARL approach: its rich environment model and its elaborate cognitive model of agents, can therefore capture aspects of socio-ecological systems which were so far overlooked or deemed too complicated to model appropriately such as the implications of needing to learn to implement one's strategic decisions.

With MARL modeling of human interactions in complex social-ecological systems it is possible for group-level social phenomena to emerge from the interaction of individual agents' cognitive models, without direct specification by the researcher (Figure 3). This setting allows decisions not to be made in an 'orderly fashion' with each player submitting a single strategic decision per round like in a matrix game Janssen et al. (2010). While the complexity of the environment and number of parameters in the agents make formal mathematical analysis impossible, it offers many experimental advantages. Experimental setups similar to computer games allow for rich data collection and require sophisticated analysis (Mobbs et al., 2021). Additionally, experimental conditions are easy to control, repeatable, and agents can be subjected to behavioral experiments during which they do not learn (Köster et al., 2022). This allows researchers to directly test a diverse range of hypotheses. It captures in a single framework numerous cognitive hypotheses, e.g. those about innate cognitive modules or biases, and various environmental hypotheses (e.g. concerning conditions favoring the emergence of territoriality) (McKee et al., 2021). In the following, we enumerate other consequences of specifying rich environments where open-ended learning by MARL agents is possible.

3 MARL for social-ecological systems

3.1 Modeling individual decision making

There are several ways in which simulated agents learning via interaction with a complex environment resemble humans learning by trial and error in social-ecological systems. First, agents gradually learn about the world around them, acquiring skills, which enable more elaborate actions and strategies. Their learning process is constrained by their experience and the stochasticity of the environment. Second, cognitive abilities, such as working memory and attention, may limit the fidelity and amount of information they retain over time. In addition, perceptual biases constrain the way they sample their environment, and the information available for learning and direction of future actions. These have implications for the emergent behaviour of agents in different settings.

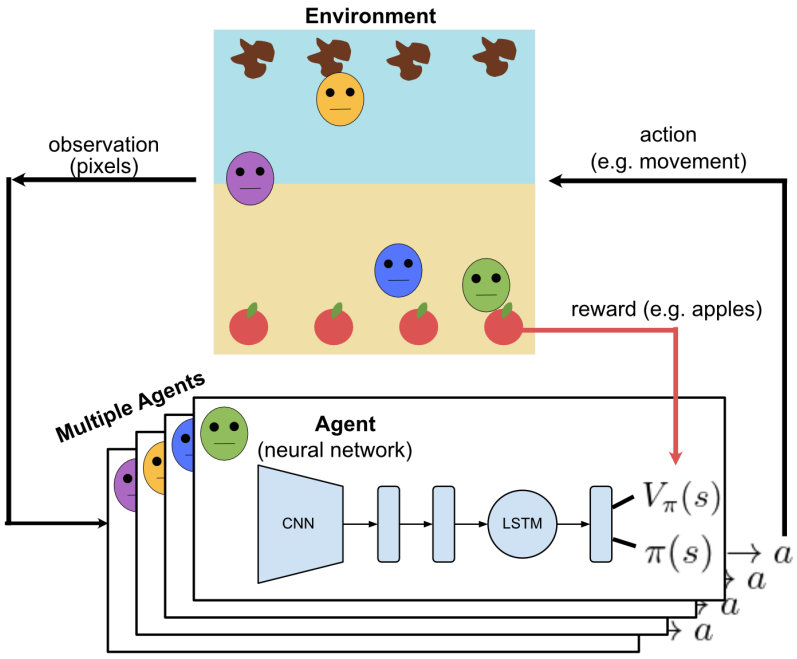

Figure 3: MARL procedure and architecture. Agents learn together while playing the game together in the same game environment. Each agent is modeled using an independent neural network. The input to the agent's model is observation of the game state in form of raw pixels, which includes the location of agents and rewards. The model can also receive social observations of other agents' actions, either by setting up such parameters explicitly or adding some visual marker for behavior that is delivered in the raw-pixel input. Each agent's (in this case an actor-critic formulation) neural network's output is an updated estimation of the current state's value ($V(s)$), and the next action a to be taken, which is encoded in a policy ($\pi(s)$) that tries to maximize the agents' reward. The agent also observe any rewards (r) that arise from the current state (s). These rewards are used to train the network and adjust its parameters, updating the state values and policies. This process takes place over many timesteps per episode, across many episodes. This allows agents to learn complex cooperative strategies to maximize their rewards of a wide horizon of states.

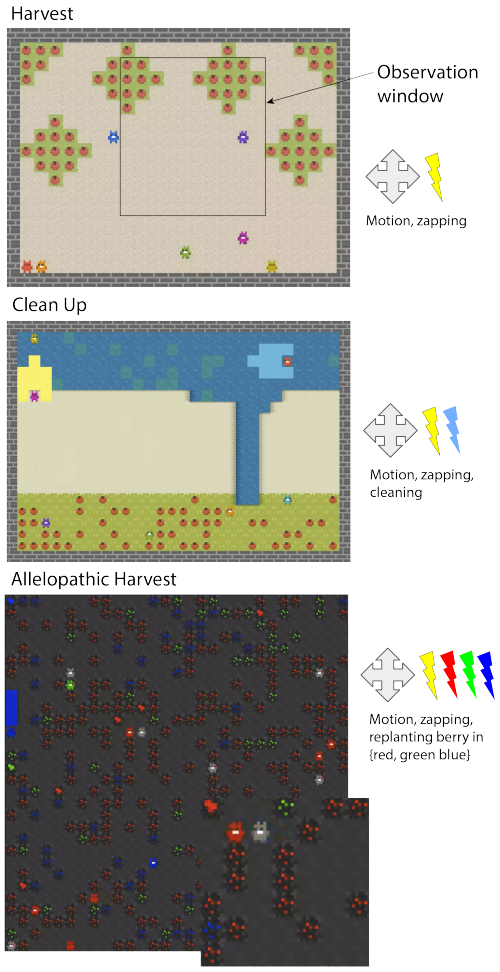

Figure 4: Three example MARL environments available via https://github.com/deepmind/meltingpot. In the top panel the observation window of the purple agent is indicated. Each environment has a different action space that includes motion, zapping, cleaning (Clean Up) and replanting berries (Allelopathic Harvest).

3.1.1 Progressive skill learning

In a social-ecological system, in order to pursue a high level strategy like cooperation, agents need to implement complex sequences of actions. In fact, cooperation could require different actions every time, since it always needs to be implemented in a different environment state. This can only be achieved if agents learn generalizable notions of the environment state and appropriate actions to take in it. This shifts the focus to learning skills and highlights how agents' cognitive limitations may hinder or help said learning.

As agents alter their environment, they can affect their own learning and the learning of others. Typically, agents first need to learn how to visually parse the world and what affordances the simulation offers them. Once these basic competencies are in place the task they face changes. For example, early in learning an agent's most pressing issue may be how to find and consume rewarding fruits. Once all agents have learned how to consume fruits, the most pressing issue they face may concern navigating competition for fruit with other agents. This creates an open-ended dynamic in which the population of agents keeps creating new and harder problems for itself (B. Baker et al., 2020; Jaderberg et al., 2019; Leibo, Hughes, et al., 2019; Plappert et al., 2021). Climbing this tree of skills in itself poses a hard learning problem. It requires the ability to continually add skills without forgetting what one has learned before, or returning to random behavior between learning different skills. For example, in a game where a bartering economy emerged, this first relied on agents learning to farm fruit for their own consumption, before they could learn that farming the fruit they are good at farming in order to trade it for other fruit they prefer to eat can be a more efficient way to obtain their favorite kind of fruit, and finally learning that, since others trade, there is an opportunity to specialize in a merchant-like behavior which entirely forgoes farming and instead seeks to profit from spatially heterogeneous prices (Johanson et al., 2022). All these behaviors depend not just on one's own experience but also on the skills and behaviour of others, a factor which changes over time and thus may constantly provide new learning opportunities.

3.1.2 Cognitive constraints

Like humans, but unlike many game theoretic models, MARL agents are typically constrained in their pursuit of reward by their cognitive abilities and knowledge about the world (Simon, 1990). For example, an agent may have a working memory module with limited capacity and time horizon. Working memory capacity has been shown in artificial agents to affect the emergence of group-based discrimination behavior in a setting where statistical discrimination may emerge on the basis of agents inappropriately conditioning partner choice behavior on features that only spuriously correlate with partner quality. Agents with more working memory capacity are better able to ignore spurious correlations and thereby learn less discriminatory policies (Duéñez Guzmán et al., 2021). Similarly, spatial attention capacity is another limiting factor. If

agents' fields of view are typically restricted, they may neglect to explore sufficiently to discover far away rewarding locations (Leibo, Perolat, et al., 2019). The most important cognitive restriction is that agents need to discover and learn the mechanics of the world. These considerations generally mirror human cognition in which decision making is constantly constrained both by cognitive capacities and knowledge of the environment (Simon, 1990).

Notably, since basic perceptual learning is modeled endogenously in MARL, it is possible for relatively low-level perceptual biases to arise. For example, in situations where larger and brighter objects are relevant for reward, they produce larger gradients and thus faster learning relative to alternative situations where smaller and duller objects are reward-relevant (Leibo et al., 2018) simply because they are associated with larger numbers (higher RGB values). This can mean that, on a population level, which convention emerges may be influenced by which option initially appears more salient. For instance, in 'Allelopathic Harvest' (see Fig. 4), groups of agents were more likely to converge on conventions featuring the replanting of brightly-colored berry varieties (over dull-colored varieties) (Köster et al., 2020). Another line of work exploring the downstream effects of perceptual biases in multi-agent reinforcement learning concerns the emergence of statistical discrimination (stereotyping). In this model agents learn discriminatory partner selection policies (i.e. selecting partners by pixel color) when they fail to distinguish spurious correlations from the truly causal features of their social environment. This effect, driven by perceptual ease, can be ameliorated by increased computational power of each agent, allowing them to learn to perform less biased and more accurate partner selection (Duéñez Guzmán et al., 2021). Such salience effects arise because artificial agents include a connectionist model of basic visual perception which is subject to some similar idiosyncrasies as biological vision systems (Leibo et al., 2018; Lindsay, 2021). These effects can give rise to 'focal points' and are thus of importance as natural default options in the study of convention formation (Schelling, 1960).

For humans, a wide array of perceptual and cognitive biases may be related to perception of others' actions, as in the case of omission bias (Ritov & Baron, 1995), and in the processing of punishments and rewards (Palminteri & Pessiglione, 2017). When learning about action-outcome contingencies one learns about the likelihood of an action to lead to a specific outcome. When outcomes are caused by an active action, such as an agent zapping another agent and harming them, this is relatively easy. However, when an outcome is caused by an omission act, e.g., when an agent does not move an obstacle, thus keeping another agents' path to reward blocked, people find it hard to associate the outcome with the omission act (Ritov & Baron, 1995), therefore attenuating the learning process. Such biases can accumulate in multi-agent environments, causing different collective behaviours to emerge and persist. For example, when human players in a multiplayer foraging game learn about social norms that govern the other players behaviour, they learn faster about norms that are manifested in active rather than passive actions, and norms that entail harmful rather than beneficial outcomes (Hertz, 2021).

3.2 Environmental constraints

Social-ecological systems include an environment, whose spatial and temporal characteristics can influence which types of collective behaviors emerge, e.g. territoriality or turn-taking. In addition, the way rewards are distributed among players, e.g. whether rewards are publicly available or privately held, can dramatically shape the emergent collective behaviors and social structures.

3.2.1 Spatial characteristics

Experimenters can set the environment's spatial characteristics, making rewards either concentrated in one part of space or uniformly distributed, and adding spatial constraints on movement. For example, uniform physical space can become part of a social compromise. In a common pool resource problem, an environment filled with fruits that cannot regrow once they are fully harvested, territoriality emerged as part of a more sustainable solution. Agents equipped with a zapper learned to deter other agents from coming too close, creating effectively privatized spaces. While it was possible to harvest these spaces more sustainably for the individual agent, this also increased inequality across agents (Perolat et al., 2017). In experiments with human participants, when punishment is not possible but participants are allowed to communicate, they divide up the space in equal size regions (Janssen, 2010; Janssen et al., 2010).

Emergent territoriality preferences are akin to peripersonal space in humans. Peripersonal space refers to the space immediately surrounding our body, usually defined as the space which we can reach and manipulate objects (Serino, 2019). Studies show that people prefer that other people wouldn't enter their peripersonal space, which they keep as a buffer around them for defensive purposes (Graziano & Cooke, 2006). The amount of peripersonal space that people prefer to keep clear around them is context dependent, as we may allow some people to be closer to us some of the time, and it is also varied across individuals, as the amount of space needed increases with social anxiety levels, for example (Givon-Benjio et al., 2020).

3.2.2 Temporal Characteristics

As simulations unfold over time, and strategic policies may include sequences of many discrete actions, they allow the emergent of complex temporal dynamics. For instance, temporally-coordinated turn-taking behavior may emerge in the Clean-Up environment (Figure 1). In this environment agents collect apples in one location, but there is also a river on the opposite side of the environment which they need to maintain in a healthy state by periodically traveling there to clean pollution which gradually accumulates. Apples cannot grow when the amount of pollution in the river gets too high (Hughes et al., 2018). Without coordination, all agents might eat until there are no more apples, and then all go to clean at the same time, all individually motivated by the dearth of apples. This creates equal outcomes, but is inefficient. With more coordination, a subset of agents could clean more often than the others, who are then free

to focus on eating. This is efficient since it reduces travel time but it creates unequal outcomes. An emergent solution that has equal outcomes and increased efficiency is when players all asynchronously alternate between cleaning and eating according to non-overlapping schedules.

Such rotation systems are also common in human social-ecological systems, e.g. taking turns using irrigation systems, where and when to fish, and coordinating where and when to graze and provide manure (Ostrom et al., 1994). In laboratory experiments with human subjects we see that participants rotate locations when some areas have higher growth than other locations (Janssen, 2010), or when participants have asymmetric access to a shared resource like in irrigation systems (Janssen et al., 2011). Studying such complex arrangements spanning the interaction of space and time would not be possible in matrix games since they assume actions are submitted in lockstep (e.g. two players deciding whether to cooperate or defect in the same timestep). The spatial and temporal resolution of environments like Clean Up enables the study of how extended strategies are implemented step by step (McKee et al., 2021).

3.2.3 Situations with multiple equilibria and heterogeneous tastes

Agents often prefer to align their behavior with that of others. When environments feature multiple equilibria each one is called a "convention". E.g. driving on the left or on the right side of the road are both workable equilibria, all agents prefer to align to whichever solution the others around them have also selected. For humans, entrenched conventions often take on normative force, prescribing the behavior of group members (Ullmann-Margalit, 2015). These conventions, which may also be called social norms, are maintained by conformity of group members (Hertz, 2021), and by enforcement and punishment of members that do not comply (Fehr & Schurtenberger, 2018; Heyes, 2022; Sripada & Stich, 2006).

Norms and conventions allow humans to coordinate and cooperate without direct communication. This can be done by simulating the other party's decision process, and predict what their actions are going to be based on current settings, norms, and our beliefs about the other's motivations (Misyak & Chater, 2014). This process was suggested to include a Bayesian inference process, in which one's beliefs about other's intentions are updated online as the other's actions unfold (C. L. Baker et al., 2009). Indeed, endowing agents with the ability to negotiate norms and coordinate behavior based on such internal models seems to be an important step in creating social artificial agents that can operate in human society (Chater, 2023). However, this inference process is cognitively demanding, and while humans are able to perform deep strategic inference, they avoid such deep planning in favor of shallower inference and heuristics , in a way that balances between cognitive resource investment and outcome, e.g., material payoff, maximization (Levine et al., n.d.; Lieder & Griffiths, 2019). It is therefore important to take such consideration into account when designing artificial agents that are suppose to model human behavior and interact with humans, for example by adjusting their cost function (Alanqary et al., 2021;

Fränken et al., 2024).

One environment where convention formation has been studied with MARL is called 'Allelopathic Harvest' (Köster et al., 2020). In this environment, agents like to eat berries, of which there are three varieties (red, green, and blue). They grow in a fixed number of berry patches, each of which can be replanted to grow any variety. Each variety suppresses the growth rate of the others, so berries grow faster when the environment state is closer to monoculture. At any time, agents can replant berries of any variety in any patch. This setup features both a "start-up problem" and a free-rider problem. The start-up problem refers to the difficult of coordinating agents to plant the same variety as one another when the current state is far from monoculture. Lack of coordination discourages investment (planting) since agents do not want to "back the wrong horse" Marwell & Oliver (1993). The free-rider problem refers to the choice which arises after some progress has been made toward monoculture between spending additional time contributing to the common good by planting versus focusing instead on gathering berries for oneself.

The Allelopathic Harvest setting is further complicated by the fact that agents have heterogeneous tastes. Some agents receive intrinsically more reward from consuming a particular berry variety over the others (e.g. more reward from consuming red berries than green or blue berries). Agents disagree with one another as to which convention would be optimal, a situation called normative disagreement (Stastny et al., 2021). The strength of the emergent convention can be measured by the fractions of berries of each variety, i.e. a high fraction of red berries represent a high realization of the red 'convention', that is established berry by berry. The worst outcome for the group is a total lack of coordination, yielding the lowest rewards for everyone. As all agents get some reward for all berries, any convention is better for every single agent than no convention at all. This is akin to a system like a parliamentary democracy, where multiple conflicting goals can be pursued and each individual can choose which goal to support, and face continual temptation to free ride on the efforts of others. Despite only being equipped with model-free learning, artificial agents are able to form conventions in this environment. After time, some agents explicitly start to support a convention that has traction, despite it being against their intrinsic taste preference (Köster et al., 2020).

3.2.4 Communication

Complex environments where agents may communicate with one another contain a range of challenges which do not arise in abstracted matrix games settings. Communication channels may be either ungrounded and symbolic or grounded in the affordances admitted by the underlying simulated environment Dor (2023).

In the ungrounded case there is a large literature on the spontaneous emergence of communication between reinforcement learning agents. One study showed that emergent communication patterns in abstracted reference games may be grounded in natural language by co-training networks on both an in-

teractive task (a multi-agent referential game) and a passive task (supervised image-labeling task with natural language image labels) (Lazaridou et al., 2017). More recent work in this area considered zero-shot coordination protocols where agents must rapidly adapt to new partners who may have learned different "languages". In this setting agents require some form of regularization to encourage their use of non-arbitrary symbols, which can be reliably used to coordinate with unfamiliar co-players as long as they experienced a similar environment and independently settled on a similar vocabulary of signals, an effect that may be useful for modeling the quasi-universality of gesture-based communication (Bullard et al., 2020; Hu et al., 2020; Zhu et al., 2021).

In the case of grounded communication channels, several papers studied agents endowed with actions to negotiate over the exchange of objects. These were instances of grounded communication since the trades would automatically resolve when agreed, transferring the objects between players' inventories (Johanson et al., 2022; Zheng et al., 2022). Various properties of supply and demand emerged in these studies. However, they had a substantive limitation in that agents were able to automatically maintain private property in their inventory and exchange simultaneously without risk of theft, making the problem of learning to trade easier than it would otherwise be. However, a later study extended this line of work to show the emergence of trade and tolerated theft using much more generic 'pick up' and 'drop' actions (Garbus & Pollack, 2023). Similarly, research in experimental economics demonstrated that exchange of goods between human participants only happens if property rights are defined (Kimbrough et al., 2008).

3.3 Cooperation in social dilemmas

When agents are assumed to be fully self-interested, both traditional game theory and MARL predict that agents should not cooperate in social dilemma situations. These prediction are at odds with human behavior in real life as well as in the laboratory (Camerer, 2011). Both game theory and MARL must be augmented in some way to account for the data on human cooperation, trust, and altruism.

The approach to studying cooperation in MARL models has mirrored the approach in behavioral game theory. In both cases the research workflow starts by observing that the theory wrongly predicts self-interested individuals should not be able to solve a problem (e.g. find a way to cooperate). It then proceeds to modify the basic model to encode social preferences, and finally to show that the augmentations are sufficient inducement for agents to find cooperative solutions, either through rational decision-making (in behavioral game theory) or learning (in MARL).

3.3.1 Augmenting utility functions (intrinsic motivation)

This approach weakens the assumption that agents are purely self-interested by adding terms to agent utility functions encoding various kinds of "social prefer-

ences" like altruism. We already mentioned Fehr & Schmidt (1999) inequality avoidance bias, which modeled human behavior in simple economic games by introducing utility function terms reflecting an aversion to inequity causing them to prefer certain money allocations over others Fehr & Schmidt (1999a). Following Fehr & Schmidt (1999), a MARL paper adapted the inequity aversion approach to MARL and showed it remained effective at promoting cooperation in much more complex multiagent settings (Hughes et al., 2018). In this case the aversion signal was interpreted as coming from within the agent itself, as an 'intrinsic motivation' (Chentanez et al., 2004). Players in this game typically receive rewards for eating fruits. However, the intrinsic motivation term also subtracts small amounts of reward depending on how much reward all other players earned recently. Notably, being averse to advantageous inequity allows more favourable outcomes in the public-goods-like clean up game since agents were less interested in eating while others did the work of cleaning, and thus explored cleaning actions more often. Disadvantageous inequity aversion on the other hand had a different effect. It encouraged agents to punish those who take more than others, and thereby indirectly creates a situation where overharvesting is discouraged via punishment. As a result, disadvantageous inequity aversion promoted sustainable behavior in a common-pool resource appropriation game (Hughes et al., 2018).

Another way to account for human sensitivity to others' well-being is through the construct of social value orientation, a cognitive construct representing how an individual trades off one's own gain against the gain of others (Griesinger & Livingston Jr, 1973). It can be seen as a personality factor which reflects an individual's general inclination toward altruism versus self-interest. In computational models it is possible to endow artificial agents with social value orientation by making them sensitive not only to their own rewards but also to the rewards of other agents in the population (Li et al., 2023; McKee et al., 2020). Unsurprisingly, in mixed motive games (e.g. the aforementioned Clean Up), groups with overall more altruistic social value orientation achieve higher and more equal group outcomes. A surprising result is that in many cases, heterogeneous groups fare better than groups with the same mean (but homogeneous) social value orientation (B. Baker, 2020; McKee et al., 2020).

3.3.2 Reciprocity

In iterated Prisoner's Dilemma, agents become incentivized to cooperate if their partner always punishes defection by defecting themselves (tit-for-tat) (Axelrod, 1984). Likewise many laboratory experiments have shown that humans are *conditional cooperators*—i.e. they cooperate as long as others also cooperate Fischbacher et al. (2001). Tit-for-tat-like conditional cooperation strategies are also effective at promoting cooperation in complex MARL environments. One approach uses hierarchical MARL agents where a hard-coded high-level controller implementing tit-for-tat decides whether to play a cooperating policy, trained with joint reward, or a defecting policy, trained using the default self-interested rewards (Kleiman-Weiner et al., 2016; Lerer & Peysakhovich, 2017).

This approach yields cooperation in complex spatially and temporally extended settings via reciprocity.

Sometimes reciprocity can be achieved by augmenting utility functions. In behavioral game theory, Rabin (1993) took such an approach to modeling conditional cooperation in matrix games, proposing a model where agents represent the "kindness" of their actions and the perceived kindness of others' actions. This model augments the agent's utility function to include terms encouraging the agent to match the kindness of their own actions to the perceived kindness of the others (Rabin, 1993). Eccles et al. (2019) took an analogous approach for MARL in complex environments. Agents in this model achieve cooperation via reciprocity because they learn to recognize the "niceness" level of others' behavior and then imitate their niceness level back to them. The motivation to imitate is encoded by an augmentation to their utility function that sets a value on imitation. This approach performed well in several different spatiotemporally complex environments including Clean Up (Eccles et al., 2019).

3.3.3 Reputation

The social-ecological systems we described are complex and provide many low-level observations that one could keep track off, and many events happen outside one's field of view. Multiple agents-side mechanism described here provide agents with useful summary statistics of the environment. Reputation is a specific summary statistic which tracks social behavior over time, and is seen as an important contributor to the ability to sustain cooperation (Milinski et al., 2002). In McKee et al. (2021), agents were equipped with 'reputation' information in the game of Clean Up. They were able to see who cleaned how much, and they had an intrinsic aversion to deviating from the group (in either direction). Agents equipped with this consideration for reputation, achieved better group results in Clean Up. Importantly, this study directly drew parallels between agent and human behavior. Humans in this task too, achieved better group outcomes when reputation information was presented to them. In particular, both artificial agents and humans achieved cooperation in a very similar way—they used reputation information to help increase turn taking behavior.

Another intrinsic social motivation is not directly related to rewards that other agents achieve, but rather to the actions other agents take. Having people voluntarily seeking and following an individual's advice is one manifestation of prestige and social status (Cheng et al., 2010). A number of studies indicate that humans adapt their behaviour in order to influence other people's behaviour (Hertz et al., 2017, 2020; Lindström et al., 2021). While agents who only seek to influence others may produce deceptive behaviour in some scenarios (Kurvers et al., 2021; Schwardmann & van der Weele, 2019), humans seem to give accurate advice and promote their accuracy as advisers, i.e. they tend to act prosocially as a way to build social prestige (Atkisson et al., 2012; Hertz et al., 2020; Zaatri et al., 2022). Thus individual desire to gain social influence can promote collective benefit and cooperation. In order to capture such results, one MARL study looked at agents who estimated their social influence over

others and received an intrinsic reward when they took highly influential actions. In Clean Up, using these internally generated predictions of other agents' actions to internally reward oneself for influential actions helped groups discover cooperative equilibria (Jaques et al., 2019).

3.3.4 Norms

Both game theory and MARL have employed a variety of different concepts of 'norm', 'social norm', and the like. The various accounts differ in part because they are aimed at different levels of analysis. They appear to divide into two conceptually distinct categories: (a) norms as utility transformations, and (b) norms as equilibria (Grossi et al., 2013). However, the very same norm can usually be seen through either lens. For example, consider a norm that proscribes a certain action. From the individual's point of view, the effect of this norm will be felt like an augmentation to a utility function (reducing the payoff of the proscribed action). Yet at the same time, we may consider how the rule itself arises from aggregate behavior of all the individuals comprising the group (Guala, 2016). Both levels of analysis are important in MARL, with some studies focusing on how individuals learn to enforce and comply with norms (Köster et al., 2022), while other studies concern how the norms themselves may emerge (Vinitsky et al., 2023). Agents have also been shown to profit by being furnished with Bayesian rule induction that allows them to rapidly infer which norms are being followed by a specific population (Oldenburg & Zhi-Xuan, 2024).

For example, several researchers studied how cooperation may be achieved in complex social dilemmas by endowing agents with the ability to directly modify one another's rewards, e.g. by sending positive reward as inducement to act prosocially (Lupu & Precup, 2020; W. Z. Wang et al., 2021) or via negotiated contracts for future reward-transfer commitments (Christoffersen et al., 2022; Willis & Luck, 2023). Both levels of analysis of the norm concept are evident simultaneously in this line of work. The incentives faced by each individual arise from the pattern of inducements offered by their coplayers. And at the same time, the pattern of gifting and contracting behaviors themselves evolve as agents learn.

In a study where individuals learned to enforce and comply with norms, agents had to gather berries while avoiding one poisonous berry variety (Köster et al., 2022). It was prohibitively difficult to identify the poisonous berry by trial and error since agents would rapidly eat berries of all varieties, and the time delay of the poisonous effect was long enough to make it impossible to accurately identify which specific berry was the cause of the deleterious effect. However, when the environment 'marked' those who ate the poisonous berry and rewarded others for punishing marked agents, all quickly learned to avoid the poisonous berry. Unsurprisingly, swifter punishment made this credit assignment problem more tractable. Counterintuitively, this result held even if the introduced extra rule was a 'silly rule' which proscribes the eating of a perfectly harmless berry. Since the negative effects of the actually poisonous berry were so detrimental it was worth it for the group to engage in "wasteful" punish-

ments for consuming the harmless "taboo" berry just for the extra opportunity to practice enforcement and compliance that they created (Köster et al., 2022). The study's main result was to show that there were benefits from introducing additional rules, that provide more practice for agents to learn effective punishment and compliance behavior. This kind of result could only be achieved because all agent actions were constituted of the reusable building blocks of low-level actions orchestrated by the cognitive machinery of the agent that allows for generalization.

4 Open problems and future directions in MARL models of social-ecological systems

In this paper we presented a new experimental approach, built on MARL, to study complex social-ecological systems. The MARL approach allows examining of how spatial and temporal characteristics of such systems shape collective behavior. We demonstrated the flexibility of this approach, and how it can be augmented to incorporate additional environmental affordances, motivational structures, and other mechanisms in order to address social-ecological characteristics, cognitive processes, social structures, and dynamics. And, we discussed how the constant interplay of these factors makes possible that a whole range of group-level phenomena can emerge without being directly specified. We conclude by highlighting some open questions that demonstrate the potential of coupling this modeling approach with studies of real-world settings. Such studies of human social cognition may facilitate identifying the tolerances or working ranges where environments lead to emergent behaviours, versus points where the emergent behaviour changes, and thus provide a powerful tool to formulate novel scientific insights and predictions of real-world dynamics.

4.1 MARL as a discovery research tool

In this work we surveyed work that used the MARL approach to capture collective behavior demonstrated in humans, and the way insights from research of human cognition can be used to enhance these models. Both lines of work together provide a foundation for new studies where MARL can be used to provide novel insights about factors shaping human behavior in social-ecological systems. In much of the MARL work discussed above, MARL was used to examine which factors of the environment were important to give rise to human-like behavior. To get MARL systems to replicate human phenomena, the researchers needed to set up specific cognitive abilities and environmental settings. This approach allows understanding the physical and cognitive factors that may underlie human behavior, and provides testable predictions. It is possible to expand on it by formally comparing agents with different cognitive models, different motivations, and different environmental affordances to test specific hypotheses concerning the way humans form collective decisions and cooperation patterns. For example, it is possible to test how social motivations and abilities associated with

different levels of social anxiety support and stabilize social structure (Brosnan et al., 2017; Zaatri et al., 2022), or how environmental affordances can provide a unified framework for coordination (Small & Adler, 2019). Some works already close the loop, testing whether the behavioral patterns predicted by MARL simulations hold in human behavior(McKee et al., 2021). Others, such as the study on spurious normativity (Köster et al., 2022), provide novel explanations to human phenomena. We hope that our work demonstrates MARL's potential as a research tool, and will inspire social and cognitive scientists to incorporate it in their research.

It is important to also be mindful of the mismatches between MARL and humans. While typically deep RL agents learn gradually, humans excel at rapid acquisition of knowledge (like episodic memory) or rapid inference or reasoning taking advantage of a model of the world. An artificial agent approaching a task as a blank slate, starting with a random neural network, stands in stark contrast to a human brain formed by the course of evolution and a lifetime of individual experience. For instance, the often sparse environmental rewards in MARL typically do not attempt to account for the many metabolic needs that biology balances when making decisions. However, in principle such detail can be introduced into the MARL framework to yield more realistic simulations (e.g. Ackley & Littman (1991)).

4.2 Better models for the effect of communication

When human participants in collective action experiments can communicate, even without the ability to enforce promises, cooperation improves significantly (Janssen et al., 2010). However, at present there are no MARL models explaining why communication should have this effect. Why does it work in humans? There are several possible mechanisms: communication may allow coordination among participants, they may develop trust relationships, they may express social pressure, etc, but there is no conclusive explanation of the effect of communication in commons dilemma experiments (DeCaro et al., 2021). In fact, it seems that the nature of communication (e.g. constructive) is more important than the specific content of the communication (DeCaro et al., 2021). Future work using AI methods could explore these hypotheses.

4.3 Predicting the effect of interventions

Sufficiently rich social-ecological system simulations may be used to plan interventions in a way that could be tailored to specific human problems and capacity. In one early example of this approach, researchers constructed a 2-D world containing mechanics that allow for a simple economy (resources, buildings) populated by four artificial agents. A fifth artificial agent was tasked with setting a tax policy for this virtual world, guiding the population to a state that was both more prosperous and more equal (Hua et al., 2023). This approach could be further enhanced by attempting to quantitatively model human economic behavior and preferences using recurrent neural networks. This approach

directly injects data gathered from human behavior into the simulation. It stands in contrast to the "first-principles" approach described above, in which the aim was to show a qualitative agreement between the pattern of results produced by the model and those obtained from a group of human participants.

One useful technique for quantitatively fitting human behavior is called 'Behavioral cloning'. It has been used successfully to increase artificial agent performance on specific games by incorporating examples of how humans play them (B. Baker et al., 2022; Carroll et al., n.d.). This technique can be used to create virtual players to populate a simulation, and can be done in a way that generalizes to new situations not experienced in the training data. For example, it was possible to design harder psychological tasks using behavioral clones, and that then these tasks were indeed found later to be more difficult by human players (Dezfouli et al., 2020). Using these methods, it may eventually be possible to construct complex socio-economical system models using artificial agents whose behavior is grounded in human data (Koster et al., 2022). Such models may ultimately connect ecological complexity and human-like decision-making, a juxtaposition which will be needed to explore consequences of climate change on human behavior and to simulate the effectiveness of environmental policies (Schill et al., 2019).

sectionAcknowledgement UH was supported by the Israel Science Foundation (1532/20).

5 Declaration of interests

The authors declare no competing interests.

References

Ackley, D., & Littman, M. (1991). Interactions between learning and evolution. *Artificial life II*, *10*, 487–509.

Agapiou, J. P., Vezhnevets, A. S., Duéñez-Guzmán, E. A., Matyas, J., Mao, Y., Sunehag, P., ... others (2022). Melting pot 2.0. *arXiv preprint arXiv:2211.13746*.

Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020). Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, *117*(16), 8825-8835. doi: 10.1073/pnas.1915841117

Alanqary, A., Lin, G. Z., Le, J., Zhi-Xuan, T., Mansinghka, V. K., & Tenenbaum, J. B. (2021, June). Modeling the mistakes of boundedly rational agents within a bayesian theory of mind. *arXiv [cs.AI]*.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Organizational influence processes*, *58*, 295–303.

Atkisson, C., O'Brien, M. J., & Mesoudi, A. (2012). Adult learners in a novel environment use prestige-biased social learning. *Evol. Psychol.*, *10*(3), 519–537. doi: 10.1177/147470491201000309

Axelrod, R. M. (1984). *The evolution of cooperation*. Basic Books.

Baker, B. (2020). Emergent reciprocity and team formation from randomized uncertain social preferences. *Advances in Neural Information Processing Systems*, *33*, 15786–15799.

Baker, B., Akkaya, I., Zhokhov, P., Huizinga, J., Tang, J., Ecoffet, A., ... Clune, J. (2022). Video pretraining (VPT): Learning to act by watching unlabeled online videos. *arXiv preprint arXiv:2206.11795*.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2020). Emergent tool use from multi-agent autocurricula. *International Conference on Learning Representations*.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., & Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science*, *359*(6373), 325–329.

Biggs, R., De Vos, A., Preiser, R., Clements, H., Maciejewski, K., & Schlüter, M. (2021). *The routledge handbook of research methods for social-ecological systems*. Taylor & Francis.

Blair, G., Littman, R., & Paluck, E. L. (2019). Motivating the adoption of new community-minded behaviors: An empirical test in nigeria. *Science advances*, *5*(3), eaau5175.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron*, *107*(4), 603–616.

Brosnan, S. F., Tone, E. B., & Williams, L. (2017). The evolution of social anxiety. In T. K. Shackelford & V. Zeigler-Hill (Eds.), *The evolution of psychopathology* (pp. 93–116). Cham: Springer International Publishing. doi: 10.1007/978-3-319-60576-0_4

Bullard, K., Meier, F., Kiela, D., Pineau, J., & Foerster, J. (2020). Exploring zero-shot emergent communication in embodied multi-agent populations. *arXiv preprint arXiv:2010.15896*.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton university press.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2018, September). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat Hum Behav*, *2*(9), 637–644. doi: 10.1038/s41562-018-0399-z

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., & Dragan, A. (n.d.). On the utility of learning about humans for human-AI coordination. *Advances in neural information processing systems*.

Chater, N. (2023, July). How could we make a social robot? a virtual bargaining approach. *Philos. Trans. A Math. Phys. Eng. Sci.*, *381*(2251), 20220040. doi: 10.1098/rsta.2022.0040

Cheng, J. T., Tracy, J. L., & Henrich, J. (2010). Pride, personality, and the evolutionary foundations of human social status. *Evol. Hum. Behav.*, *31*(5), 334–347. doi: 10.1016/j.evolhumbehav.2010.02.004

Chentanez, N., Barto, A., & Singh, S. (2004). Intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, *17*.

Christoffersen, P. J., Haupt, A. A., & Hadfield-Menell, D. (2022). Get it in writing: Formal contracts mitigate social dilemmas in multi-agent rl. *arXiv preprint arXiv:2208.10469*.

DeCaro, D. A., Janssen, M. A., & Lee, A. (2021). Motivational foundations of communication, voluntary cooperation, and self-governance in a common-pool resource dilemma. *Current Research in Ecological and Social Psychology*, *2*, 100016.

Dezfouli, A., Nock, R., & Dayan, P. (2020). Adversarial vulnerabilities of human decision-making. *Proceedings of the National Academy of Sciences*, *117*(46), 29221–29228.

Donges, J. F., Heitzig, J., Barfuss, W., Wiedermann, M., Kassel, J. A., Kittel, T., ... Lucht, W. (2020). Earth system modeling with endogenous and dynamic human societies: the copan: Core open world–earth modeling framework. *Earth System Dynamics*, *11*(2), 395–413.

Dor, D. (2023, March). Communication for collaborative computation: two major transitions in human evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *378*(1872), 20210404. doi: 10.1098/rstb.2021.0404

Du, Y., Leibo, J. Z., Islam, U., Willis, R., & Sunehag, P. (2023). A review of cooperation in multi-agent learning. *arXiv preprint arXiv:2312.05162*.

Duéñez Guzmán, E. A., McKee, K. R., Mao, Y., Coppin, B., Chiappa, S., Vezhnevets, A. S., ... Leibo, J. Z. (2021). Statistical discrimination in learning agents. *arXiv:2110.11404 [cs.LG]*.

Eccles, T., Hughes, E., Kramár, J., Wheelwright, S., & Leibo, J. Z. (2019). Learning reciprocity in complex sequential social dilemmas. *arXiv preprint arXiv:1903.08082*.

Elsawah, S., Pierce, S. A., Hamilton, S. H., Van Delden, H., Haase, D., Elmahdi, A., & Jakeman, A. J. (2017). An overview of the system dynamics process for integrated modelling of socio-ecological systems: Lessons on good modelling practice from five case studies. *Environmental Modelling & Software*, *93*, 127–145.

Fehr, E., & Schmidt, K. M. (1999a). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, *114*(3), 817–868.

Fehr, E., & Schmidt, K. M. (1999b, August). A theory of fairness, competition, and cooperation. *Q. J. Econ.*, *114*(3), 817–868. doi: 10.1162/003355399556151

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*(7), 458–468. doi: 10.1038/s41562-018-0385-5

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404.

Folke, C., Pritchard Jr, L., Berkes, F., Colding, J., & Svedin, U. (2007). The problem of fit between ecosystems and institutions: ten years later. *Ecology and society*, *12*(1).

Freundlieb, M., Kovács, Á. M., & Sebanz, N. (2016). When do humans spontaneously adopt another's visuospatial perspective? *J. Exp. Psychol. Hum. Percept. Perform.*, *42*(3), 401–412. doi: 10.1037/xhp0000153

Fränken, J.-P., Valentin, S., Lucas, C. G., & Bramley, N. R. (2024, January). Naïve information aggregation in human social learning. *Cognition*, *242*(105633), 105633. doi: 10.1016/j.cognition.2023.105633

Garbus, J., & Pollack, J. (2023). Emergent resource exchange and tolerated theft behavior using multi-agent reinforcement learning. *arXiv preprint arXiv:2307.01862*.

Givon-Benjio, N., Oren-Yagoda, R., Aderka, I. M., & Okon-Singer, H. (2020). Biased distance estimation in social anxiety disorder: A new avenue for understanding avoidance behavior. *Depression and Anxiety*, *37*(12), 1243–1252.

Graziano, M. S., & Cooke, D. F. (2006). Parieto-frontal interactions, personal space, and defensive behavior. *Neuropsychologia*, *44*(6), 845–859.

Griesinger, D. W., & Livingston Jr, J. W. (1973). Toward a model of interpersonal motivation in experimental games. *Behavioral science*, *18*(3), 173–188.

Grossi, D., Tummolini, L., & Turrini, P. (2013). Norms in game theory. *Agreement Technologies*, 191–197.

Guala, F. (2016). Understanding institutions. In *Understanding institutions.* Princeton University Press.

Henrich, B. J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & Mcelreath, R. (2001). In search of homo economicus : Behavioral experiments in 15 Small-Scale societies. *American Economic Review*, *91*(2).

Hertz, U. (2021, June). Learning how to behave: cognitive learning processes account for asymmetries in adaptation to social norms. *Proc. Biol. Sci.*, *288*(1952), 20210293. doi: 10.1098/rspb.2021.0293

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., & Bahrami, B. (2017, December). Neural computations underpinning the strategic management of influence in advice giving. *Nat. Commun.*, *8*(1), 2191. doi: 10.1038/s41467-017-02314-5

Hertz, U., Tyropoulou, E., Traberg, C., & Bahrami, B. (2020, October). Self-competence increases the willingness to pay for social influence. *Sci. Rep.*, *10*(1), 17813. doi: 10.1038/s41598-020-74857-5

Heyes, C. (2022, June). *Rethinking norm psychology.* doi: 10.31234/osf.io/t5ew7

Hoffman, M., & Yoeli, E. (2022). *Hidden games: the surprising power of game theory to explain irrational human behaviour.* Basic Books.

Hu, H., Lerer, A., Peysakhovich, A., & Foerster, J. (2020). "other-play" for zero-shot coordination. In *International conference on machine learning* (pp. 4399–4410).

Hua, Y., Gao, S., Li, W., Jin, B., Wang, X., & Zha, H. (2023). Learning optimal "pigovian tax" in sequential social dilemmas. *arXiv preprint arXiv:2305.06227*.

Hughes, E., Leibo, J. Z., Philips, M. G., Tuyls, K., Duéñez-Guzmán, E. A., Castañeda, A. G., ... Graepel, T. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in neural information processing systems* (pp. 3330–3340).

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castaneda, A. G., ... others (2019). Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, *364*(6443), 859–865.

Janssen, M. A. (2010). Introducing ecological dynamics into common-pool resource experiments. *Ecology and Society*, *15*(2).

Janssen, M. A., Holahan, R., Lee, A., & Ostrom, E. (2010). Lab experiments for the study of social-ecological systems. *Science*, *328*(5978), 613–617.

Janssen, M. A., M., A. J., & Joshi, S. (2011). Coordination and cooperation in asymmetric commons dilemmas. *Experimental Economics*, *14*(4), 547-566.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., ... De Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning* (pp. 3040–3049).

Johanson, M. B., Hughes, E., Timbers, F., & Leibo, J. Z. (2022). Emergent bartering behaviour in multi-agent reinforcement learning. *arXiv preprint arXiv:2205.06760*.

Kimbrough, E. O., Smith, V. L., & Wilson, B. J. (2008). Historical property rights, sociality, and the emergence of impersonal exchange in long-distance trade. *American Economic Review*, *98*(3), 1009-1039.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *Cogsci.*

Koster, R., Balaguer, J., Tacchetti, A., Weinstein, A., Zhu, T., Hauser, O., ... others (2022). Human-centred mechanism design with democratic AI. *Nature Human Behaviour*, *6*(10), 1398–1407.

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*, *119*(3), e2106028118.

Köster, R., McKee, K. R., Everett, R., Weidinger, L., Isaac, W. S., Hughes, E., ... Leibo, J. Z. (2020). Model-free conventions in multi-agent reinforcement learning with heterogeneous preferences. *arXiv preprint arXiv:2010.09054*.

Kurvers, R. H. J. M., Hertz, U., Karpus, J., Balode, M. P., Jayles, B., Binmore, K., & Bahrami, B. (2021). Strategic disinformation outperforms honesty in competition for social influence. *iScience*, *24*(12), 103505. doi: 10.1016/j.isci.2021.103505

Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. In *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings.*

Leibo, J. Z., d'Autume, C. d. M., Zoran, D., Amos, D., Beattie, C., Anderson, K., ... others (2018). Psychlab: a psychology laboratory for deep reinforcement learning agents. *arXiv preprint arXiv:1801.08116*.

Leibo, J. Z., Dueñez-Guzman, E. A., Vezhnevets, A., Agapiou, J. P., Sunehag, P., Koster, R., ... Graepel, T. (2021). Scalable evaluation of multi-agent reinforcement learning with Melting Pot. In *International conference on machine learning* (pp. 6187–6199).

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 preprint arXiv:1903.00742*.

Leibo, J. Z., Perolat, J., Hughes, E., Wheelwright, S., Marblestone, A. H., Duéñez-Guzmán, E., ... Graepel, T. (2019). Malthusian reinforcement learning. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems* (pp. 1099–1107).

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th international conference on autonomous agents and multiagent systems (aa-mas 2017)*. Sao Paulo, Brazil.

Lerer, A., & Peysakhovich, A. (2017). Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*.

Levin, S., Xepapadeas, T., Crépin, A.-S., Norberg, J., De Zeeuw, A., Folke, C., ... others (2013). Social-ecological systems as complex adaptive systems: modeling and policy implications. *Environment and development economics*, *18*(2), 111–132.

Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. (n.d.). Resource-rational contractualism: A triple theory of moral cognition. *osf.io*.

Li, W., Wang, X., Jin, B., Lu, J., & Zha, H. (2023). Learning roles with emergent social value orientations. *arXiv preprint arXiv:2301.13812*.

Lieder, F., & Griffiths, T. L. (2019, February). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.*, *43*(e1), e1. doi: 10.1017/s0140525x1900061x

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, *33*(10), 2017–2031.

Lindström, B., Bellander, M., Schultner, D. T., Chang, A., Tobler, P. N., & Amodio, D. M. (2021, December). A computational reward learning account of social media engagement. *Nat. Commun.*, *12*(1), 1311. doi: 10.1038/s41467-020-19607-x

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, *30*.

Lupu, A., & Precup, D. (2020). Gifting in multi-agent reinforcement learning. In *Proceedings of the 19th international conference on autonomous agents and multiagent systems* (pp. 789–797).

Léger, C., Hamon, G., Nisioti, E., Hinaut, X., & Moulin-Frier, C. (2023, December). Evolving reservoirs for meta reinforcement learning. *arXiv [cs.LG]*.

Marwell, G., & Oliver, P. (1993). *The critical mass in collective action*. Cambridge University Press.

McKee, K. R., Gemp, I., McWilliams, B., Duèñez-Guzmán, E. A., Hughes, E., & Leibo, J. Z. (2020). Social diversity and social preferences in mixed-motive reinforcement learning. In *Proceedings of the 19th international conference on autonomous agents and multiagent systems* (pp. 869–877).

McKee, K. R., Hughes, E., Zhu, T. O., Chadwick, M. J., Koster, R., Castaneda, A. G., ... Leibo, J. Z. (2021). A multi-agent reinforcement learning model of reputation and cooperation in human groups. *arXiv preprint arXiv:2103.04982*.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of abnormal and social psychology*, *67*(4), 371.

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002). Reputation helps solve the 'tragedy of the commons'. *Nature*, *415*(6870), 424–426.

Misyak, J. B., & Chater, N. (2014). Virtual bargaining: a theory of social decision-making. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *369*, 20130487. doi: 10.1098/rstb.2013.0487

Mobbs, D., Wise, T., Suthana, N., Guzmán, N., Kriegeskorte, N., & Leibo, J. Z. (2021, July). Promises and challenges of human computational ethology. *Neuron*, *109*(14), 2224–2238. doi: 10.1016/j.neuron.2021.05.021

Nisioti, E., & Moulin-Frier, C. (2020). *Grounding artificial intelligence in the origins of human behavior.* arXiv.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., ... Vazire, S. (2022, January). Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.*, *73*, 719–748. doi: 10.1146/annurev-psych-020821-114157

Oldenburg, N., & Zhi-Xuan, T. (2024). Learning and sustaining shared normative systems via bayesian rule induction in markov games. *arXiv preprint arXiv:2402.13399*.

Ostrom, E. (1998). A behavioral approach to the rational choice theory of collective action: Presidential address, american political science association, 1997. *American Political Science Review*, *92*(1), 1–22.

Ostrom, E. (2009). A general framework for analyzing sustainability of social-ecological systems. *Science*, *325*(5939), 419–422.

Ostrom, E., Gardner, R., & Walker, J. (1994). *Rules, games and common-pool resources*. University of Michigan Press.

Palminteri, S., & Pessiglione, M. (2017). Opponent brain systems for reward and punishment learning: causal evidence from drug and lesion studies in humans. In *Decision neuroscience* (pp. 291–303). Elsevier.

Perolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., & Graepel, T. (2017). A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in neural information processing systems*, *30*.

Plappert, M., Sampedro, R., Xu, T., Akkaya, I., Kosaraju, V., Welinder, P., ... Zaremba, W. (2021). Asymmetric self-play for automatic goal discovery in robotic manipulation. *arXiv preprint arXiv:2101.04882*.

Poteete, A. R., Janssen, M. A., & Ostrom, E. (2010). *Working together: collective action, the commons, and multiple methods in practice*. Princeton University Press.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American economic review*, 1281–1302.

Ritov, I., & Baron, J. (1995, November). Outcome knowledge, regret, and omission bias. *Organ. Behav. Hum. Decis. Process.*, *64*(2), 119–127. doi: 10.1006/obhd.1995.1094

Savage, L. J. (1954). *The foundations of statistics*. Courier Corporation.

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

Schill, C., Anderies, J. M., Lindahl, T., Folke, C., Polasky, S., Cárdenas, J. C., ... Schlüter, M. (2019). A more dynamic understanding of human behaviour for the anthropocene. *Nature Sustainability*, *2*(12), 1075–1082.

Schlager, E., Blomquist, W., & Tang, S. Y. (1994). Mobile flows, storage, and self-organized institutions for governing common-pool resources. *Land Economics*, 294–317.

Schwardmann, P., & van der Weele, J. (2019, October). Deception and self-deception. *Nature Human Behaviour*, *3*(10), 1055–1061. doi: 10.1038/s41562-019-0666-7

Sebanz, N., Bekkering, H., & Knoblich, G. (2006, February). Joint action: bodies and minds moving together. *Trends Cogn. Sci.*, *10*(2), 70–76. doi: 10.1016/j.tics.2005.12.009

Serino, A. (2019, April). Peripersonal space (PPS) as a multisensory interface between the individual and the environment, defining the space of the self. *Neurosci. Biobehav. Rev.*, *99*, 138–159. doi: 10.1016/j.neubiorev.2019.01.016

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... others (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, *362*(6419), 1140–1144.

Simon, H. A. (1990). Bounded rationality. In *Utility and probability* (pp. 15–18). Springer.

Small, M. L., & Adler, L. (2019, July). The role of space in the formation of social ties. *Annu. Rev. Sociol.*, *45*(1), 111–132. doi: 10.1146/annurev-soc-073018-022707

Sripada, C. S., & Stich, S. (2006). A framework for the psychology of norms. *The innate mind*, *2*, 280–301.

Stastny, J., Riché, M., Lyzhov, A., Treutlein, J., Dafoe, A., & Clifton, J. (2021). Normative disagreement as a challenge for cooperative AI. *arXiv preprint arXiv:2111.13872*.

Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., ... Vicente, R. (2017). Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, *12*(4), e0172395.

Ullmann-Margalit, E. (2015). *The emergence of norms*. Oxford University Press, USA.

Vinitsky, E., Köster, R., Agapiou, J. P., Duéñez-Guzmán, E. A., Vezhnevets, A. S., & Leibo, J. Z. (2023). A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence*, *2*(2), 26339137231162025.

Wang, J. X., Hughes, E., Fernando, C., Czarnecki, W. M., Duéñez-Guzmán, E. A., & Leibo, J. Z. (2019). Evolving intrinsic motivations for altruistic behavior. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems* (pp. 683–692).

Wang, W. Z., Beliaev, M., Bıyık, E., Lazar, D. A., Pedarsani, R., & Sadigh, D. (2021). Emergent prosociality in multi-agent games through gifting. *arXiv preprint arXiv:2105.06593*.

Willis, R., & Luck, M. (2023). Resolving social dilemmas through reward transfer commitments.

Wilson, J., Yan, L., & Wilson, C. (2007). The precursors of governance in the maine lobster fishery. *Proceedings of the National Academy of Sciences*, *104*(39), 15212–15217.

Wilson, J. A., Acheson, J. M., Metcalfe, M., & Kieban, P. (1994). Chaos, complexity and community management of fisheries. *Marine Policy*, *18*(4), 291–305.

Young, O. R. (2002). *The institutional dimensions of environmental change: fit, interplay, and scale*. MIT press.

Zaatri, S., Aderka, I. M., & Hertz, U. (2022, May). Blend in or stand out: social anxiety levels shape information-sharing strategies. *Proc. Biol. Sci.*, *289*(1975), 20220476. doi: 10.1098/rspb.2022.0476

Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., & Socher, R. (2022). The AI economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science advances*, *8*(18), eabk2607.

Zhu, H., Neubig, G., & Bisk, Y. (2021). Few-shot language coordination by modeling theory of mind. In *International conference on machine learning* (pp. 12901–12911).

Zimbardo, P. G. (1969). The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In *Nebraska symposium on motivation.*