2022

Humans in Algorithms, Algorithms in Humans: Understanding Cooperation and Creating Social AI with Causal Generative Models

Lux Miranda
University of Central Florida

HUMANS IN ALGORITHMS, ALGORITHMS IN HUMANS: UNDERSTANDING COOPERATION AND CREATING SOCIAL AI WITH CAUSAL GENERATIVE MODELS

by

LUX MIRANDA
B.S. Utah State University, 2020

A thesis submitted in partial fulfilment of the requirements
for the degree of Master of Science in Industrial Engineering
in the Department of Industrial Engineering and Management Systems
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2022

ABSTRACT

Cooperation is the hallmark human trait which has allowed us to congregate into the vast, continent-sprawling societies we live in today. Yet, the precise social, environmental, and cognitive mechanisms which enable this cooperation are not fully understood. Toward this, lucrative insights have been borne through the use of formal computational models of socio-cognitive phenomena: In simulating our own cooperative behavior, we can better deduce the exact factors which cause it. The combined knowledge of these factors and ability to computationally simulate them allows us to further two goals: First, it empowers us with the knowledge of how to modify our social systems to better human well-being and promote more sustainable, equitable, and compassionate societies. Second, the computational aspect allows us to more directly create artificial, socially competent companions—whether robotic or entirely digital—to cooperate with us in the real world in achieving the first goal. In this thesis, I contribute to the development of artificial social cognition by examining two case studies of cooperation dilemmas: a game of social team cooperation inference known as stag-hunt, and a stylized cooperative irrigation system. Specifically, I show causal, generative models encoding hypotheses on actual mechanisms in the human mind which are able to outperform the extant state-of-the-art models in both of these cases. In the second case, I show how models like this can be automatically discovered through an algorithm known as evolutionary model discovery, greatly expediting the deduction of new models in similar domains. The results have implications not only for understanding the dynamics of human teaming and irrigation systems (the humans in algorithms), but also broader human socio-cognitive mechanisms contributing to cooperation (the algorithms in humans)—all while simultaneously allowing these mechanisms to be encoded into socially competent AI.

To all whose social systems have failed to serve them.

ACKNOWLEDGMENTS

I would be nowhere without the outstanding mentorship of Dr. Ozlem Ozmen Garibay and Dr. Ivan Garibay, who I do not think I can ever thank enough for taking a chance on me.

Thanks also to Dr. Jacopo Baggio for being an excellent human being and not minding me digging up his old code to try and hack away at improving it. Thank you as well to Dr. Luis Rabelo and the rest of my committee for weathering the administrative hiccups as I learned how to schedule a thesis defense!

I would also like to properly thank Dr. Jacob Freeman for leading me here and letting me pick so many raspberries from his backyard.

Lastly, my biggest thanks goes to my amazing partner Victoria and cats Navi and O'Brien for providing the coziness that I needed to function while making this thesis a reality.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Since the start of the Holocene, above nearly everything else, it has been our ability to cooperate with each other in groups large and small that has most thoroughly permeated the human experience and led to the planet-enveloping societies which we live in today (Ostrom, 1990; Turchin, 2016). As our social world becomes ever more complex by the day, so do our problems: We now face global-scale challenges such as the climate crisis, global inequality, pandemics, and existential risks which require us to cooperate with each other at greater scale and effectiveness than we ever have before (Michelozzi and De'Donato, 2021; Bostrom, 2013). Pursuant to this, it is more critical than ever that we have a precise understanding of cooperation and its mechanisms so that we might leverage it to save ourselves and better our world.

Yet, the precise social, environmental, and cognitive mechanisms which enable and dictate the outcome of cooperative efforts—mechanisms which constantly interact with and mutually shape each other—are not fully understood (Ostrom, 1990; Ostrom et al., 1999; Turchin, 2016; Dietz et al., 2003; Baggio et al., 2015). Toward developing a formal understanding of these mechanisms, we begin with the source: The human mind. Complex social phenomena and cooperative megaprojects such as space stations and countries governing millions of people all begin with simple, individual interactions between many human minds (Epstein, 2014, 1999). It therefore stands to reason that by more precisely understanding socio-cognitive processes and how they interact between individuals, we can understand how to influence the emergent behavior at the macro-scale where our global problems reside.

Toward this, lucrative insights have been borne through the use of formal computational models of such phenomena (Epstein, 2014; Baggio et al., 2015; Gunaratne and Garibay, 2020; Gunaratne et al., 2021; Rabkina, 2020). Causal, generative models—and, more specifically, their inverse

discovery—represent a new generation of modelling techniques for not only socio-cognitive processes but of any complex process emerging from small-scale interactions (Garibay et al., 2021). In this thesis, we present two case studies of advances of this kind of model: An intent recognition model for observing *stag-hunt*, a simple multiplayer game where agents must cooperate to maximize rewards, and a model of agent interaction for a social irrigation system.

Intent recognition—the human ability to utilize social and behavioral cues to infer each other's intents, infer motivations, and predict future actions—is a central process to human social life, and governs our fundamental ability to cooperate with each other (Jara-Ettinger et al., 2020). In addition to contributing to the study of cooperation, artificial agents with greater social intelligence have wide-ranging applications from enabling the collaboration of human-AI teams (Fiore et al., 2010; Fiore and Wiltshire, 2016) to improving the effectiveness of socially assistive robots (Winkle et al., 2021). In chapter 2, we show that the Naïve Utility Calculus generative model (Jara-ettinger et al., 2016; Jara-Ettinger et al., 2020) is capable of competing with leading models in intent recognition and action prediction when observing *stag-hunt*, a simple multiplayer game where agents must infer each other's intentions to maximize rewards. Moreover, we show that the model is the first with the capacity to out-compete human observers in intent recognition after the first round of observation. The chapter concludes with a discussion on implications for the Naïve Utility Calculus and of similar generative models in general.

Small-scale irrigation systems which require cooperation from multiple users to maintain are a common feature of many small farms (Cifdaloz et al., 2010; Anderies et al., 2013). Small farms are thought to produce around a third of the global crop supply, but they are also likely to be increasingly vulnerable to changes in the spatial and temporal availability of water (Anderies et al., 2013; Janssen et al., 2012). In this context, it is key to assess the effect of the social mechanisms which promote resilience in small-scale irrigation systems and, more widely, in complex social-ecological dilemmas under changing conditions. Small-scale irrigation systems are characterized

by upstream farmers having prevailing access to a canal's resources, yet all farmers along the canal must contribute to maintaining the irrigation infrastructure. In chapter 3, to further assess the ensemble of social mechanisms promoting the resilience of irrigation systems, we build on previous work in which a stylized irrigation dilemma was simulated via a social lab experiment (Anderies et al., 2013). Studies of the data produced from this experiment modeled participants' behavior with multiple, theoretically grounded agent-based models (ABMs) (Baggio and Janssen, 2013; Janssen and Baggio, 2017). These models encode causal, human-interpretable hypotheses of decision making which generates the real-world behavior observed in the experiment. However, the accuracy of these models in fitting the experimental data is limited. Using Evolutionary Model Discovery, a recent algorithm for inverse generative social science (iGSS) (Gunaratne and Garibay, 2020), we show the ability to automatically generate a wide variety of unique new ABMs which fit the experimental data more accurately and robustly than the original, manually-constructed ABMs. To do this, we algorithmically explore the space of possible behavioral rules for agents choosing how to contribute to the maintenance of the irrigation infrastructure. We find that, in contrast to the original models, our best-performing models typically have an additional element of stochasticity and favor factors such as other-regarding preferences and perceived relative income. Given that this change in just a small part of the original model has yielded such an advance,

Our results suggest that causal generative models and iGSS methods have great potential for continuing to derive more accurate models of complex emergent phenomena, be it for social inference, the sustainability of agricultural systems, or any application beyond.

Statement of Originality

Parts of this work have been included in conference presentations and a work under review for journal publication. Other than the work discussed in the following manuscripts, the rest of this

thesis has not been published publicly at the time of writing:

- Miranda, L. and Ozmen Garibay, O. (2021). Multi-agent Naive Utility Calculus: Intent Recognition in the Stag-Hunt Game. In Thomson, R., Hussain, M. N., Dancy, C., and Pyke, A., editors, *Social, Cultural, and Behavioral Modeling*, Lecture Notes in Computer Science, pages 331–340, Cham. Springer International Publishing

- Miranda, L., Ozmen Garibay, O., and Baggio, J. (2022). Evolutionary model discovery of human behavioral factors driving decision-making in an irrigation experiment. In review, *Journal of Artificial Societies and Social Simulation*

CHAPTER 2: CASE STUDY 1: INTENT RECOGNITION IN STAG HUNT

Introduction

A fundamental process in our everyday life is our ability to observe peoples' actions, infer their beliefs, desires, and intentions, and predict what they will do next.

Consider this example: Your roommate mentions they are hungry and gets up from their seat. You infer they are headed to the kitchen. You may know from observing their past behavior that, at this time of day, they are likely to get a bowl of cereal. You may, then, decide to inform them that all of the bowls are currently unwashed. Having been informed thus, your roommate decides they are not hungry enough to bother putting forth the effort to wash a bowl, and they slump back into their chair.

Here, from just a few observations of your roommate's behavior, you made a correct inference of their desires and propagated that inference to predict their plan of future actions. You performed an intervention—mentioning the dirty bowls—and this provided them with new information on how to act on their desires. Your roommate, thus, determined that the reward gained from eating a bowl of cereal was less than the cost of washing a bowl. In other words, your roommate determined their plan of action had a *negative subjective utility*—so they chose not to act.

If one may see the great complexity of mental inferences in this simple social interaction, it is easy to appreciate the great difficulty (and powerful consequences) of creating an algorithm capable of replicating these processes. Indeed, a great many algorithms have been created, studied, and used for just this type of intent recognition in a variety of domains and contexts (Sukthankar et al., 2014; Demiris, 2007; Qi and Zhu, 2018); this chapter principally focuses on evaluating a novel computational framework (the Naïve Utility Calculus) in inferring cooperative intent among

multiple agents in a shared environment.

One application of immediate interest for this particular inference problem is its use in human-AI teaming. Algorithmic intent recognition of cooperation between humans is critical to developing AI that can enhance teams' coordination and efficacy in performing complex tasks such as urban search-and-rescue (Fiore et al., 2010; Fiore and Wiltshire, 2016; Barnes et al., 2017) or socially assist people in a variety of therapeutic and care contexts Winkle et al. (2021). Furthermore, there has been a shift in recent years in scholars beginning to view cognitive science and generative social science as interdependent fields (Orr et al., 2018), and many models have been advanced by using more complex and realistic cognitive architectures for agents (Epstein, 2014; Gunaratne et al., 2021; Baggio and Janssen, 2013; Schlüter et al., 2017). Thus, improving these cognitive architectures has consequences for, for example, detecting and preventing threats to public safety (Demiris, 2007) such as the spread of disinformation on social media (Garibay et al., 2020; Rajabi et al., 2020), or informing policy to better deal with the effects of the global climate crisis (Freeman et al., 2020; Elsawah et al., 2020).

Research in cognitive psychology indicates that the paradigm of assuming agents act to maximize utility—while not a very accurate model of actual cognitive processes—might, in fact, be a good approximation of the mind's process of inferring *other* minds' intentions (Jara-ettinger et al., 2015, 2016; Jara-Ettinger et al., 2020). Formally, such a utility function is simply defined as

$$\text{Utility}(plan, outcome) = \text{Reward}(outcome) - \text{Cost}(plan) \tag{2.1}$$

The Naïve Utility Calculus (NUC), a recently formalized framework for action-understanding, utilizes this paradigm as the basis of its function (Jara-Ettinger et al., 2020). While, prior to this work, the NUC displayed much promise as a general model for action-understanding, it had only

been put to use in a single-agent setting with limited inferences on social behavior.

Thus, this chapter is presented with the primary purpose of testing the Naïve Utility Calculus in a multi-agent setting with greater requirements on social inference. To do this, we utilize a well-studied cooperative action game known as *stag-hunt*. Stag-hunt is a multiplayer game where agents work to maximize their rewards by choosing to pursue and capture either a high-reward stag or a low-reward hare. While low-reward hares can be captured individually, agents must work together in order to capture a high-reward stag. Thus, critical to performing well in the game is the ability to infer other players' intentions to determine whether pursuing a stag is worth the effort.

While stag-hunt was introduced by Skyrms (2003) and famously used via a Minecraft implementation known as *Pig Chase* by the Microsoft Malmo Collaborative AI Challenge (Johnson et al., 2016), this work primarily draws on a version of stag-hunt used to more directly test general models of artificial theory of mind. This version, introduced by Shum et al. (2019), includes data from human subjects performing the same tasks as their computational model, which relies on Bayesian inference over a generative model encoding relations known as Composable Team Hierarchies (CTHs). Further utilized by Rabkina and Forbus (2019) with the introduction of a model known as Analogical Reasoning, this version of stag-hunt allows for a streamlined, simple experiment with directly comparable results between competing models.

We find the Naïve Utility Calculus model is, in terms of both accurately recognizing intent and predicting actions, comparable to these prior two models—with the exception of intent recognition after the first timestep. Intriguingly, after observation of this timestep, NUC significantly outperforms both prior models *and* the human subjects in intent recognition.

We begin by giving an overview of each of the three models under comparison. We describe our experimental setup in detail and the modifications required to adapt the existing NUC implementation to a multi-agent setting. Lastly, we compare results to that of the prior two models and the

human subjects.

Background

The Naïve Utility Calculus

The Naïve Utility Calculus (NUC) is a recently formalized framework for action-understanding (Jara-Ettinger et al., 2020). It operates under the assumption that agents tend to choose their actions in order to maximize some notion of subjective utility. Research in childhood psychology indicates that this paradigm is likely a good approximation of how individuals reason about the behavior of others (Jara-ettinger et al., 2015, 2016).

Within NUC, agents are treated as boundedly rational. The calculus itself is implemented as a generative model of each agents' mental processes. The model details that agents first begin by observing their environment. From this observation, the agent determines their *beliefs* about the environment and the *costs* of performing available actions. The agent's *desires* (rewards) are combined with their world belief to form possible *goals* to pursue. These goals are then considered alongside the costs of the actions needed to meet those goals, and an *intention* (plan of actions) is formed. This plan is then followed to perform actions themselves, but is subject to change should any part of the environment also change and cause the agent to re-evaluate beliefs and costs (Jara-Ettinger et al., 2020).

Observers, however, are not aware of the underlying cost and reward functions that other agents are using to decide on the actions they take. Thus, given only the actions that an agent has taken, Bayesian inference may be run over the model to produce inferred estimates of the agent's cost and reward functions. These estimated functions may then be used as input to the generative model, producing a predicted set of actions the agent will take (Jara-Ettinger et al., 2020).

In Jara-Ettinger et al. (2020), the authors introduced a formal computational model of the Naïve Utility Calculus and extensively tested it in a series of single-agent experiments. In each of these experiments, both the model and human observers were presented with images of an astronaut exploring an extra-terrestrial grid world with various types of terrain, collectible "care packages," and goal locations. The experiments found that, in this simple setting, the model was sufficient for matching the abilities of human subjects to estimate the utility function of a single agent in a variety of scenarios. However, the implications of social reasoning for these experiments are limited—only a semi-social scenario is created during a single experiment where "collectible packages" are replaced with "rescuable astronauts." However, "other agents" were merely treated as part of the environment—the same as care packages—and thus did not constitute a true multi-agent environment.

In this work, we pit the Naïve Utility Calculus up against two prior models tested in the multi-agent stag-hunt game. Analogical Reasoning, the most recent of these models, is likewise a computational formalization of qualitative theory from research in childhood psychology, but it instead formalizes the separate cognitive process of analogical thinking (Rabkina and Forbus, 2019; Rabkina, 2020). The second model is a hierarchical Bayesian model introduced by Shum et al. (2019) that explicitly encodes causal models known as Composable Team Hierarchies (CTHs) to aid in cooperation inferences.

Analogical Reasoning

The Analogical Reasoning model (Rabkina and Forbus, 2019), similar to NUC, is a computational formulation inspired by findings in cognitive psychology. It takes after the process of human reasoning known by the same name. The basic idea is that an observer learns a set of relations in one case (the *base*) and, when they encounter a previously unobserved case (the *target*), supposes

that this same set of relations holds. Hence, the observer is reasoning by analogy (Falkenhainer et al., 1989).

The primary reasoning engine used by Rabkina and Forbus (2019) is a modern iteration of an analogical reasoning algorithm known as the Structure-Mapping Engine (Forbus et al., 2017). Cases in stag-hunt are represented in the form of how the spatial relationships between agents change over each timestep in addition to several non-spatial events such as a target being captured. Analogical Reasoning is a traditional machine learning method; the authors trained and tested their model via leave-one-out cross validation over the nine scenarios (i.e., for each scenario, the model was trained on the eight other scenarios and tested on the one).

Composable Team Hierarchies

Like NUC, the Composable Team Hierarchies (CTH) model presented in Shum et al. (2019) is a generative model over which Bayesian inference may be performed. Instead of encoding the internal mental state of agents, however, the CTH model assumes that agent actions are based around a certain team hierarchy and encodes a causal model of how a given team structure would dictate agents' future actions. Bayesian inference over the model accepts observations of agent actions to infer a team hierarchy (e.g., Players A and C are teaming up against Player B), and that inferred team hierarchy is then used to predict agent actions (e.g., Players A and C will pursue a stag, Player B will pursue a hare).

Experiment

We test the abilities of the Naïve Utility Calculus using observations of the multi-agent stag-hunt game à la Rabkina and Forbus (2019) and Shum et al. (2019). Originally introduced as an alter-

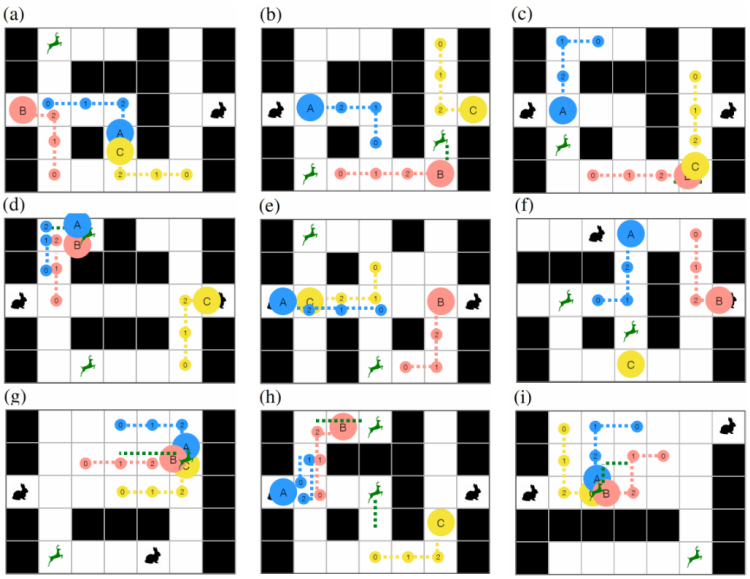

Figure 2.1: The nine scenarios of the stag-hunt game. Used with permission from Rabkina and Forbus (2019), originally adapted from Shum et al. (2019).

native to the prisoners dilemma (Skyrms, 2003), the stag-hunt presents a simple, effective, spatial schema to test inference in a variety cooperative and non-cooperative scenarios.

The game operates on the premise that a group of hunters must each choose to attempt to capture either a hare or a stag. Hares provide a low reward and can be captured by a single hunter. Stags provide a much higher reward but require a team of two or more hunters to be captured. There is no penalty for not capturing a target. Thus, it is critical for a successful hunter to determine which other hunters intend to cooperate if they wish to maximize their reward by capturing a stag.

The version of the game used here places three hunters, two stags, and two hares into a 5x7 grid world. In this grid world, there are traversable white tiles ("floors") and non-traversable black tiles ("walls"). At each timestep, hunters may move up, down, left, or right, but not diagonally. Hares are stagnant, and the movement of stags is pre-determined as part of the map structure. In general, stags attempt to move away from pursuing hunters.

This configuration is used to construct nine unique scenarios that each encode different possibilities of agent cooperation and non-cooperation. For each scenario, three timesteps of agent movement are encoded. Successful stag captures occur in five scenarios (Fig. 2.1 a, c, d, g, i), indicating cooperation, while only hares are captured in the other four scenarios (Fig. 2.1 b, e, f, h), indicating no cooperation.

Following in the steps of Shum et al. (2019) and Rabkina and Forbus (2019), these board states are used on two fronts: Firstly, the game encodes sufficient information for an observer to deduce each agent's goal by the third time step, and, consequentially, information to see which agents are cooperating to capture a stag (and which are not). Indeed, in Shum and colleagues' trials with real human observers, cooperation inferences were made correctly 100% of the time by the third timestep. The challenge lies in inferring each agent's goal before this final timestep.

Secondly is the prediction of agents' future actions. As this version of the game only runs for three timesteps, this is done in three cases: Board state predictions can be made for the second timestep and the third timestep having only observed the first timestep, and a prediction can be made for the third timestep having seen the first two timesteps. This is not a task fully encompassed by either Shum and colleagues' model or human subjects, thus our accuracy metric for this task is only compared against Rabkina and Forbus's model.

Methods

We utilize the same existing implementation of the Naïve Utility Calculus used by Jara-Ettinger et al. (2020). This implementation is in the form of a Python 2.7 package known as Bishop (`https://github.com/julianje/Bishop/`). In this study, we have adapted Bishop to use in the multi-agent setting without any modifications to the base package. This was possible

as Bishop is uniquely suited for studying this version of stag-hunt, as it is designed specifically for observations of a single agent in a grid world with various objectives that may be obtained. This was utilized in Jara-Ettinger and colleagues' experiments where an astronaut is placed in an extra-terrestrial world with various collectible care packages and observers are asked to infer the astronaut's package preference along with the cost associated with each terrain.

To adapt the usage of Bishop into a multi-agent setting, we simply encode the stag-hunt game into a format that Bishop may read and run the program on each agent individually. Of note is that, although, in theory, each player incorporates information on the other hunters to make their decisions, an outside observer of the game does not necessarily need to incorporate the state of other players into their inferences on each individual hunter. For example, if an observer is attempting to determine whether Hunter B is trying to capture a hare or a stag, they do not necessarily need to know the locations of Hunter A or Hunter C, as (at least in our nine scenarios) only Hunter B's movements are sufficient information to make the inference. Thus, in service to the limitations of the existing implementation of Bishop, a separate "map" is created for each of the three hunters between the nine scenarios, and Bishop's inference functions are run on each of these maps individually.

While previous inference engines working with stag-hunt may have explicitly encoded the meaning of walls and floors, this information is not provided to the NUC. Rather, based on the agents' movements, it is up to the NUC to determine the "cost" of traversing a floor tile versus traversing a wall tile. In the traditional formulation of the game, the wall tiles are not traversable. Thus, the NUC must determine a sufficiently, arbitrarily large cost of traversing a wall tile in order to learn this rule for itself.

Lastly, we leave all specifiable hyperparameters within Bishop at their default values. All reported results below are from 100 runs of NUC at 500 samples per run. All code and data is available via

DOI: 10.5281/zenodo.4430598.

Results

Intent recognition

In line with Shum et al. (2019) and Rabkina and Forbus (2019), our primary metric for measuring the accuracy of intent recognition is a pairwise count of which hunters are cooperating. That is, there are three predictions per scenario at each timestep: whether there is cooperation between Hunters A and B, Hunters A and C, and Hunters B and C. Model predictions are measured against the true values to produce the accuracy metric.

The NUC's predictions are summarized in Fig. 2.2. Note that the precise values for the Bayesian and Human metrics are good-faith estimates from Shum and colleagues' figures determined by Rabkina and Forbus.

We find that, after the first timestep, the NUC outperforms human subjects and all other models by a minimum of eight percentage points with a very high margin of certainty. The reason for this is uncertain; one possible hypothesis to explain this is that human observers may have less confidence in their judgement when possessing only limited information, and are therefore perhaps not achieving the maximum possible inference accuracy.

After the second timestep, NUC performs better than Analogical Reasoning but not as well as humans or the Bayesian model. After the final timestep, NUC outperforms the Bayesian model and is roughly on par with Analogical Reasoning.

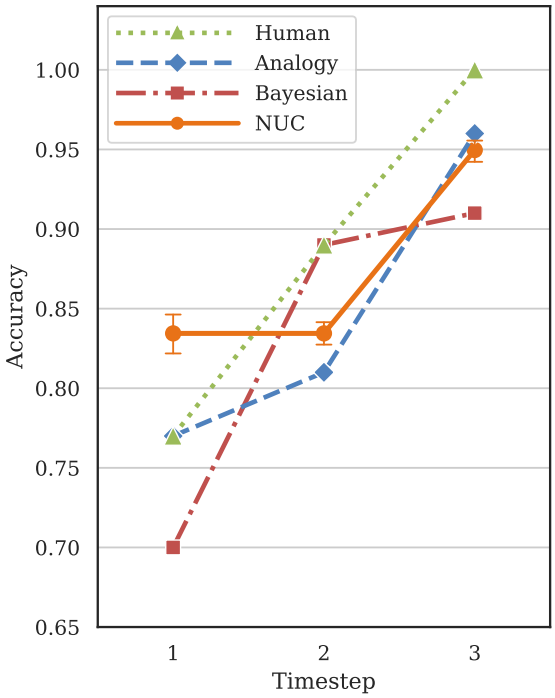

Figure 2.2: Pairwise cooperation inference accuracy across each model and the human subjects. Error bars on the NUC data indicate one standard error across 100 runs with 500 samples each. Note that accuracies for human and CTH (Bayesian) model data are based on good-faith estimates from the figures of Shum et al. (2019) and Rabkina and Forbus (2019)

Future action prediction

For the purposes of comparing NUC's action prediction ability with Rabkina and Forbus's (Rabkina and Forbus, 2019) Analogical Reasoning, we replicate their metric of measuring prediction accuracy. Due to the limitations of the current implementation of Bishop, we do not calculate action prediction accuracy in precisely the same way—while Rabkina and Forbus incorporate predicting the actions of stags into their accuracy metric, here stags are necessarily treated as parts of the environment rather than agents with predictable behavior. Thus, the metric is best used as an

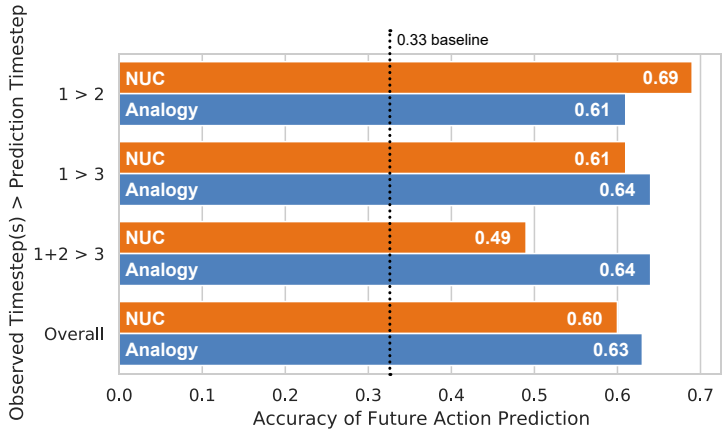

Figure 2.3: Relational action prediction accuracy across each timestep case.

approximate comparison between the two models rather than a precise one.

For each agent, NUC first performs intent recognition and then uses the inferred intent to predict its next moves. The agents' movements are measured relative to the other two agents, the two stags, and the two hares, with three possible states for the relative movements: toward, away, and stationary. For example, an action prediction for Agent B's next move might look like: Away from Agent A, Toward Agent C, Stationary to Stag 1, Toward Stag 2, Away from Hare 1, Away from Hare 2. Since there are three possible states to infer, we would expect a model guessing randomly to be correct in its prediction roughly a third of the time. Thus, the baseline accuracy for this metric is 33.33%.

From timestep 1, we make these six predictions for each agent into timestep 2 and timestep 3. Additional predictions are made given the information from timestep 2 into timestep 3. Results are summarized in Fig. 2.3. We find that, with the exception of the first timestep, NUC roughly underpreforms in comparison to Analogical Reasoning by several percentage points, though this margin might be in part due to the exclusion of stags in our prediction due to the limitations of Bishop mentioned above.

However, there is a notable case of underpreformance when predicting timestep 3 movements having observed timestep 1 and timestep 2 ($1 + 2 > 3$). We hypothesize that NUC seems to suffer from a "status quo" bias. NUC tends to fare especially poorly in cases where the first two agent movements are the same, but the agent then turns into a separate direction, forming an "L" movement pattern. We explore this quantitatively by considering raw accuracy of movement prediction on a five-fold accuracy metric predicting whether an agent will remain stationary or move up, left, down, or right. This metric, thus, has a baseline random accuracy of 20%. We find that, in the prediction of the 1+2>3 case, NUC has a raw accuracy of just 22%, with an 8% accuracy in the "L" cases and 33% accuracy in the non-"L" cases. This trend may be due to the limited nature of this version of the stag-hunt game with its small board and very few number of timesteps, but it may be a worthy line of inquiry to see if a similar pattern holds in other action-understanding problems.

Discussion

In this case study, we adapted the recently formalized computational model of Naive Utility Calculus (NUC) to the multi-agent stag-hunt game to test the model's abilities to infer agents' intent to cooperate (or not cooperate) and predict their future actions. To this end, we found the NUC's ability is comparable to leading action-understanding models which have been tested in this same version of the game. Moreover, we found that, when having only observed the first round of the game, NUC is able to outperform human observers by a significant margin in inferring which pairs of agents will cooperate.

However, the existing implementation of NUC seems to suffer from a "status-quo bias;" When attempting to predict future agent actions, the NUC strongly favored continuing established movement paths and poorly predicted sudden changes to such paths, even if such a change would bring

an agent closer to its inferred goal. It is unclear whether this is an effect due to the current implementation of NUC, NUC itself, or the limited and simple nature of the stag-hunt experiment. We advise that further research with NUC take note of this effect to see if it persists across different action-understanding scenarios.

Nonetheless, our results indicate that the Naive Utility Calculus is worthy of continued study. Particularly in light of the (albeit limited) capacity to outperform humans in some cases, the principles on which NUC is based may represent a potential avenue for advancing platforms for artificial intent recognition.

We may further interpret these results through the lens of inverse generative social science (iGSS). In the world of generative, agent-based models, it is increasingly apparent many problems are solved with not one single model which fits the problem best, but rather a class of similar models of comparable performance (Vu et al., 2019; Gunaratne et al., 2020; Gunaratne and Garibay, 2020). What we see here—three different models, none of which are clearly the "best" in all cases—may be an instance of this. Each of these models take an underlying hypothesis about how the human mind performs intent recognition and encodes it into a specific generative model. The lack, thus far, of any model's superiority may indicate that that each model may only partially detail the actual rules which the human mind is using.

In continuing the search for an intent recognition model that will match or exceed human ability, one may certainly continue to manually encode generative models based on hypotheses inspired by psychological research. These models would be both explainable and clearly motivated by real processes in the human mind. We recognize, however, that this process is very slow and ultimately bounded by the current state of psychological research. We propose, therefore, the inverse approach: computationally discovering possible models of human intent recognition, selecting the best-performers, and comparing and contrasting them to actual processes of the human mind. A

variety of emerging iGSS algorithms (e.g. Gunaratne and Garibay (2020) or Vu et al. (2019)) would be well-suited for this, though many efforts outside of the strict iGSS perspective are also underway to apply the principles of causal model discovery to machine learning more generally (Scholkopf et al., 2021).

It is our view that, with such methods, continued work may be more fruitful in discovering a more general, fundamental, causal model for social intent recognition and reasoning. Indeed, we hypothesize that the processes driving inference and decision-making in stag-hunt are not necessarily different in a fundamental way to the processes used in scenarios as diverse as, for example, reasoning in shared-resource dilemmas (Baggio and Janssen, 2013; Schlüter et al., 2017; Janssen et al., 2020), deciding how much alcohol to drink and when (Vu et al., 2019), or, even, over long time-scales across many individuals, ultimately dictating the cultural characteristics which develop on the level of entire societies (Ortman, 2018; Miranda and Freeman, 2020).

CHAPTER 3: CASE STUDY 2: COOPERATION IN IRRIGATION SYSTEMS

Introduction

As the climate crisis continues to cause ecological upset across our Earth (Michelozzi and De'Donato, 2021), an increasing number of communities have grown ever more vulnerable to once-rare environmental events such as droughts, changing temperatures, storms, and extreme weather (Michelozzi and De'Donato, 2021). Political, technological, and social changes are required in order to lessen the human cost of climate change and prevent collapse in the agricultural systems on which we all rely on. The study of social-ecological commons dilemmas, therefore, has grown more important than ever as we work to study the mechanisms which will allow us to best understand–and therefore intervene–in these systems. Increasing understanding on this front allows us to engineer and govern sustainable, antifragile systems (Taleb, 2012): systems that not only can cope, but thrive under changing conditions and increased disturbances.

These mechanisms driving human behavior in such commons dilemmas have been a focus of scholars for decades (Dawes, 1980; Ostrom, 1990; Dietz et al., 2003; Ostrom et al., 1999; Cifdaloz et al., 2010; Gutiérrez et al., 2011; Anderies et al., 2013, 2011; Janssen et al., 2012; Janssen and Baggio, 2017; Baggio et al., 2015), yet the precise mechanisms governing peoples' behavior in many types of systems are still largely a mystery. Irrigation systems are one such commons dilemma; In river and canal-based irrigation systems, upstream farmers have greater access and control of the system's resources than downstream farmers by the simple nature of their physical location. Yet, contributions are required from all users of the system in order to maintain and repair the irrigation infrastructure (e.g., Cifdaloz et al. (2010)). To date, many qualitative understandings have

been derived from the study of such systems, but no unified model exists describing the factors contributing to human decision making which are more likely to increase the resilience of these systems.

We contribute further to the study of irrigation systems by reanalyzing data collected from participants engaging in a simulated irrigation system in an experimental laboratory setting (Anderies et al., 2013). In this experiment, participants were given charge of virtual fields which required regular watering via an irrigation canal shared by three other participants. The simulation was formulated in a round-based game-like format. Participants were rewarded tokens for successfully supplying their fields with water; however, the canal's efficiency would also degrade each round, requiring participants to invest their earnings into the canal's upkeep in order to keep the system in working order.

The original experimenters gathered time series through collecting information on the participants' behavior each round such as investment amounts and water extraction levels. They deduced various macro-scale metrics from the participants' behavior such as Gini coefficients for participants' earnings and average group level investments over the entire game. Traditional statistical analysis was able to gather some mechanistic insights on the data (Anderies et al., 2013; Janssen et al., 2015; Baggio et al., 2015), but it was unable to fully explain the dynamics of each time series. Pursuant to this, scholars constructed agent-based models in attempt to better explain behavioral characteristics of the individuals in these irrigation common pool resource games (Baggio and Janssen, 2013; Janssen and Baggio, 2017). In these models, agents are modelled as participants engaging in the same round-based irrigation simulation. The studies formulated various rulesets that (stochastically) govern agents' interaction between themselves and with the resource (i.e. water and canals in an irrigation system). Each hypothesized model of agent behavior was carefully formulated based on existing theories on human decision making in relevant environments (Janssen and Baggio, 2017; Baggio and Janssen, 2013), in line with the current best practice of drawing

from cognitive science to inform agent architectures for generative social science (Orr et al., 2018; Miranda and Ozmen Garibay, 2021). However, no single model offered an especially accurate or robust explanation of the data. Each model was, at best, a partial fit.

Inverse generative social science

Recently, problems such as this have become a ripe opportunity to employ a related family of advancements in agent-based modelling known as inverse generative social science (iGSS). To conceptualize iGSS, consider: On a fundamental level, an ABM is a tacit hypothesis that a given set of micro-scale rules for agent interaction work together to produce an observed macro-scale emergent phenomena. For example, the Schelling Segregation model (Schelling, 1969) encodes the hypothesis that individuals' small preferences in the racial composition of their neighbors on the micro-scale ultimately interact to create the macro-scale emergent phenomenon of racially segregated neighborhoods. When this stylized fact is applied to real-world situations, it is, of course, just one possible hypothesis of the mechanism which may be contributing to segregated neighborhoods. While the explanation is plausible, it may only be a single facet of the entire set of dynamics driving segregation. In some cases, it may not even be at play at all (such as, in an extreme example, an authoritarian state actively enforcing segregation upon citizens with an adverse preference for it). That is, a single macro-phenomena can often be explained, or partially explained, by entirely disparate micro-scale agent rule sets.

Thus, it becomes desirable to test multiple different rulesets. Traditionally, modellers' intuitions and qualitative domain literature have been the primary sources for determining these micro-scale rules. Testing variations in the rule sets has primarily been limited to varying numeric parameters or testing only a limited selection of rule sets, based on theoretical arguments. Researchers have therefore been very limited in the number of alternative rule set hypotheses that they are able to

test, and the space of plausible, alternative rule sets goes largely unexplored. Even in the presence of simulated macro phenomena which match the observed real phenomena, there is little in the way of guaranteeing that the hypothesis chosen by the modeller is the correct process by which the phenomenon is actually generated.

Inverse generative social science (iGSS) serves to remedy this issue by leveraging machine learning to automate the exploration of the space of possible rule sets (Epstein, 1999; Vu et al., 2019). Rather than building up entire models, the iGSS modeller begins by specifying possible *subsets* of models. That is, the modeller defines a set of primitive agent constituents and operator functions capable of combining them. They then define an appropriate fitness function or selection pressure used to be able to evaluate how good a given model is, typically seeking the generation of a known macro-scale phenomenon or a precise fit to real-world data. Then, a specialized iGSS machine learning algorithm combines and recombines the model primitives into new models, evaluates them, and converges towards optimal models in a typical optimization fashion. To do this, we utilize evolutionary model discovery (EMD), a recent framework for iGSS (Gunaratne and Garibay, 2017; Gunaratne et al., 2021; Gunaratne, 2019). Our results take the form of a variety of different agent token-investment strategies and an analysis of the possible causal factors which contribute to this behavior.

In this study, we leverage advances in iGSS and evolutionary model discovery in order to devise which sets of rules are more likely to give rise to the outcomes observed in the original irrigation experiments. Our results indicate that a missing key to previously formulated models may have been an additional element of stochasticity which was not accounted for. Our analysis also indicates that other factors, such as utility maximization and perceived relative income, are also important in increasing model fitness.

In light of this and the evidence presented from other works, we argue that iGSS has powerful

potential for deriving more accurate and robust inferences on the relationship between micro-level elements and observed macro-phenomena.

Background

The irrigation experiment

This study has its origins in the tradition of studying irrigation systems. In 2010-2012, a series of experiments were run at Arizona State University with undergraduate student participants that were presented either a digitally or paper-based simulated irrigation system (Anderies et al., 2013; Janssen et al., 2015; Baggio et al., 2015). The irrigation system requires maintenance and provides water. Participants then, at each round, needed to decide how much to invest in maintaining the infrastructure and how much water they wanted to extract. Depending on the level of the group investment, a specific amount of water was available for extraction. Simulating the irrigation system implies that water extraction follows specific positions assigned to the participants, hence, the participant in position A decided how much water to extract, and then B could only extract what A left them. This procedure continued until participant E, who could only extract the water left by participants A, B, C and D. Participant order was determined before the initial round of the simulated irrigation system and remained constant over the course of the experiment. At the end of the game, the five participants were rewarded with a direct conversion of tokens to US dollars.

Multiple studies followed these experiments in order to shed light on specific individual and group characteristics in relation to the group behavior over the course of the experiment. The experiments were first statistically analyzed via regression models which discovered some correlations in the data (Anderies et al., 2013; Janssen et al., 2015). Data from irrigation experiments were also analyzed by constructing multiple, competing agent based models in which agents follow rules

dictated by theoretical considerations (i.e. selfish, altruistic, utilitarian, random etc.) (Baggio and Janssen, 2013). Further, this approach was refined on another set of irrigation experiments in Janssen and Baggio (2017). The analysis via agent based models in Baggio and Janssen (2013) and Janssen and Baggio (2017) shed some light on potential rules governing the overall irrigation system, but no model was found to clearly outperform all others. Further, these works did not test a comprehensive set of alternative rules, and nor did they test whether multiple rulesets acted at any given time.

Evolutionary Model Discovery

Here, we therefore leverage evolutionary model discovery (EMD), which utilizes machine learning to automate the discovery of causal factors driving agent decisions in agent-based models (Gunaratne and Garibay, 2020) in order to assess multiple alternative rulesets. In contrast to traditional agent-based modeling, EMD does not require that hypothesized models be manually formulated by domain experts. Instead, modelers must simply specify a set of possible factors and operators that may or may not contribute to the emergence of a phenomenon in question. They must additionally specify a fitness function that evaluates the goodness of a given model.

The algorithm on which EMD is based on combines the provided factors and operators into syntactic trees which encode new models of agent behavior. Using genetic programming, generations of these trees are formulated, mutated, and evaluated according to the provided fitness function. The best-performing rules are then allowed to cross-breed and generate successive generations. In other words, EMD combines and recombine these factors and explores the entire space of possible models before finally settling on optimal solutions (i.e. where fitness is maximized or minimized).

After obtaining the data encoding model performance from the genetic program, the modeler may use it to analyze the importance and efficacy of individual factors and their interactions. Using this

information, the modeller may analyze the automatically-generated models and use the insights gained therein to manually create new models and test whether they improve the overall fitness. These models can then be assessed in order to provide new scientific insights about the causal relationships which lead to observed macro-scale phenomena.

EMD and related approaches in iGSS have been successfully used in automating the discovery of sophisticated models in domains such as archaeology (Gunaratne and Garibay, 2020), social media analysis (Gunaratne et al., 2020), and public health (Vu et al., 2019).

Methods

Here, we specifically examine the irrigation experiment performed and analyzed in Baggio et al. (2015) and Baggio and Janssen (2013). We begin with the highest-performing model derived in Baggio and Janssen (2013): The other-regarding preferences (utilitarian) model. This model is based on findings from behavioral economics and is the most complex of the original models. In a nutshell, each agent is imbued with either a competitive, egalitarian, or altruistic disposition. The probability of an agent having any one of these decisions is determined by parameters. Each agent then incorporates information on the environment and their neighbors' behavior to make decisions which maximize some notion of utility which is congruent with their disposition.

In order to place more precise bounds on our scope and search-space, we concentrate on agent investment behavior and leave extraction behavior unchanged. Allowing the extraction behavior to also vary is a potential direction for future work. In place of the investment behavior, we allow the insertion of new models generated from a set of hypothesized alternate factors.

Hypothesized alternate factors influencing investment decision

We utilize the paradigm introduced by Agent_Zero (Epstein, 2014) of capturing a more realistic space of human behavior by ensuring factors represent three dimensions of human decision-making: rational, social, and emotional. We indicate the hypothesized factors in Table 3.1. Of these, F_{self}, F_{heur}, F_{pseu}, F_{alt}, and F_{util} represent factors originally hypothesized in Baggio and Janssen (2013). We introduce the factors: F_{rand} to serve both as a "null" model and allow additional stochasticity; F_{up} and F_{down} to allow for dynamics more directly related to neighbors' investments; and F_{inc} to add an emotional dimension, as prior work has shown that perceived income relative to others affects investment behavior (Anderies et al., 2013; Janssen et al., 2012; Baggio et al., 2015).

We formulate each factor as a function $F_i(x) : [0..10] \rightarrow [0, 1]$. This represents the probability of investing x tokens (an integer between 0 and 10) due to the given factor F_i. These factors can then be combined using addition (+), subtraction (-), multiplication (*), and division (/) into a combined rule $R(x)$. For example, a given $R(x)$ may be $R(x) = 2 * F_{\text{self}} + F_{\text{up}} - F_{\text{down}}$.

Agents then decide the number of tokens to invest, x', using argmax over the possible investment amounts:

$$x' = \underset{x \in [1..10]}{\text{argmax}}\ R(x) \tag{3.1}$$

Or, alternatively, to allow for the probabilistic complement of $R(x)$, argmin may also be used:

$$x' = \underset{x \in [1..10]}{\text{argmin}}\ R(x) \tag{3.2}$$

Note also that, with the inclusion of the subtraction operator, the probabilistic complement of each individual factor is also encoded. For example, while F_{up} produces a more positive value for x

more similar to the upstream neighbor's last investment, $-F_{\text{up}}$ produces a more negative value.

We model factors concerned with higher probability centered on a particular value v with the linear probability density

$$P(v, x) = -\left|\frac{v - x}{\omega}\right| + 1 \tag{3.3}$$

Where ω is a parameter controlling the "width" of the probabilization.

Evolutionary model discovery

Our function fit for evaluating model fitness is identical to that used in Baggio and Janssen (2013). It is defined as:

$$fit = fit_1 \cdot fit_2 \cdot fit_3 \cdot fit_4 \cdot fit_5 \tag{3.4}$$

Where each fit_i is defined as the normalized squared difference between simulated data, d_s, and experimentally observed data, d_e, for a particular metric:

$$fit_i = 1 - (d_e - d_e)^2 \tag{3.5}$$

With the following five metrics, representing five of the most charactaristic time series of the data collected:

- Average group level investments in the public infrastructure level over the 10 rounds (fit_1)
- The average contribution per position (fit_2)

- The average collection per position (fit_3)
- The average Gini coefficient of contributions (fit_4)
- The average Gini coefficient of collected tokens (fit_5)

Hence, higher fit represents a better fit to the experimental data with a maximum fitness of $fit = 1$. This fitness function is useful, in that it effectively achieves a single aggregated function encompassing multiple optimization objectives. Through multiplying each metric, low fitness in one metric is more heavily penalized than if the metrics were simply summed.

Janssen et al. (2012) compared different configurations of fitness measures (including the fit defined here), but ultimately did not find a qualitative difference in them. Hence, we choose this fit for its quantitative advantages described above.

Although we only allow the investment behavior of agents to vary in our evolutionary model discovery, we justify keeping extraction fitness as part of fit by the fact that the amount of water available for extraction is still affected by the infrastructure efficiency (which is affected by investment behavior). Thus, the extraction time series is still affected by investment behavior, and we desire a close fit to it. Moreover, as all agents use the same extraction strategy, the baseline fitness granted by this behavior is the same for all models and does not affect the comparison in overall fitness determined by variations in investment behavior.

We initialize all model parameters to a random value uniformly distributed between plus-or-minus 0.05 the optimal values reported in Table 3 of Baggio and Janssen (2013). This initialization allows us to alleviate over-fitting to arbitrarily precise parameters which allows the discovery of more robust models Gunaratne and Garibay (2020).

We chose the hyperparameters for evolutionary model discovery based on Gunaratne and Garibay

(2020): We set the mutation rate to 0.2, and we allowed the crossover rate to vary between 0.6 and 0.8. We set the minimum tree depth to 2, as this is the minimum depth for a valid rule in the formulation of the problem (allowing for the first layer to be the argmax/argmin function and the second layer to be a single factor). To encourage the exploration of the possibility of more complex rules, we set the maximum tree depth to 64.

In total, we ran the algorithm until it had evolved approximately 10,000 models. Of these, based on fitness, we selected the top 100 models and sampled each of them 100 times to obtain a fitness distribution under the randomly initialized parameters. For all of this, we used the NetLogo/Python implementation of EMD provided by Gunaratne and Garibay (2020).

Results

We begin showcasing the results by assessing all 10,000 evolved models/rulesets. Our analysis follows the methodological footsteps put forth by prior EMD analyses such as Gunaratne et al. (2020) and Gunaratne et al. (2021). That is, we begin by analyzing the importance of individual and joint factors in terms of their impact on the fitness of each model. We then analyze the best-fitness models themselves. This allows us to paint a larger picture of not only the best models, but the general significance and robustness of including factors in any given model. This allows us to more precisely see what contributes to a given models success, and in theory also allows the manual construction of new models based on insights garnered from the algorithmic construction.

The factor analysis is important for more precisely deducing the dynamics of each factor, as evolutionary model discovery was successful in evolving a great variety of rules with varying complexity in factor interaction. The set of evolved rulesets range from from rules with just one or two factors to the most complex model evolved, argmin$((F_{rand} + F_{alt} - F_{rand} - F_{up}) * F_{down}) * (((F_{heur} +$

$F_{self}) * (F_{self} + F_{heur})) + (F_{heur} * F_{heur}))$, incorporating eleven factor presences (although its fitness is only 0.218).

In order to determine the importance of individual factors, we report the results of a random forest regression with 394 trees (Figure 3.1). Both statistical dispersion (Gini) and permutation accuracy importance metrics indicate that F_{rand} is the most important factor in predicting fitness, followed by F_{util}.

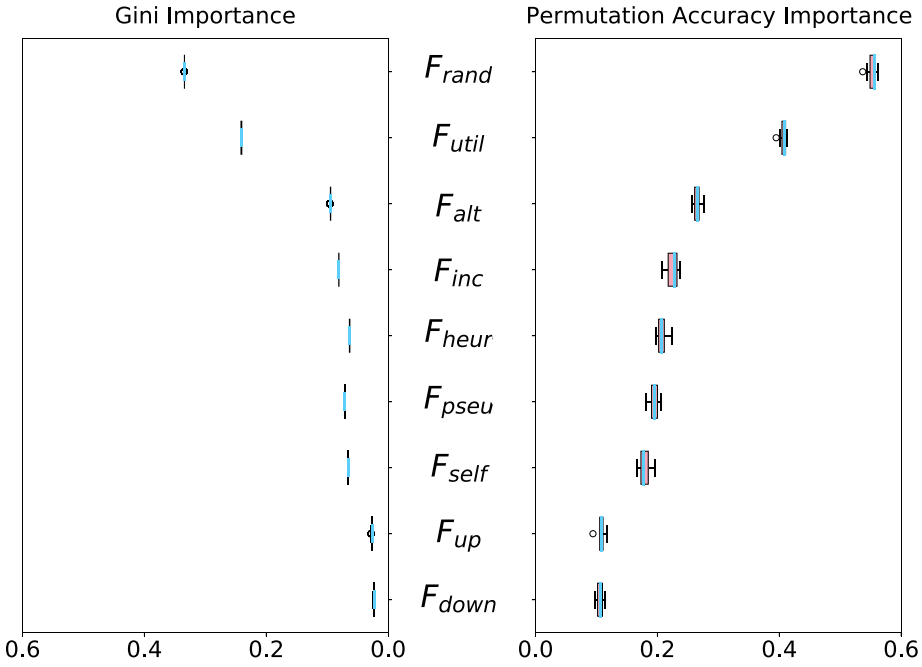

Figure 3.1: Gini importance and permutation accuracy importance of the hypothesized factors towards a random forest's ability to predict the models' fitness. F_{rand} and F_{util} display the highest Gini and permutation accuracy importance.

To further assess the factors using this importance information, we conduct a pairwise analysis comparing the importance of each factor with every other factor. Figure 3.2 shows the p-values of one-tailed Mann-Whitney U tests ($\alpha = 0.05$) comparing the permutation importance of each factor

A against every other factor B with H_0 : importance of A = importance of B and H_1 : importance of A ¿ importance of B. At least 7 of the 9 factors show significant difference and can be ordered from highest to lowest permutation accuracy importance as: $F_{\text{rand}}, F_{\text{util}}, F_{\text{alt}}, F_{\text{inc}}, F_{\text{heur}}, F_{\text{pseu}}, F_{\text{self}}$.

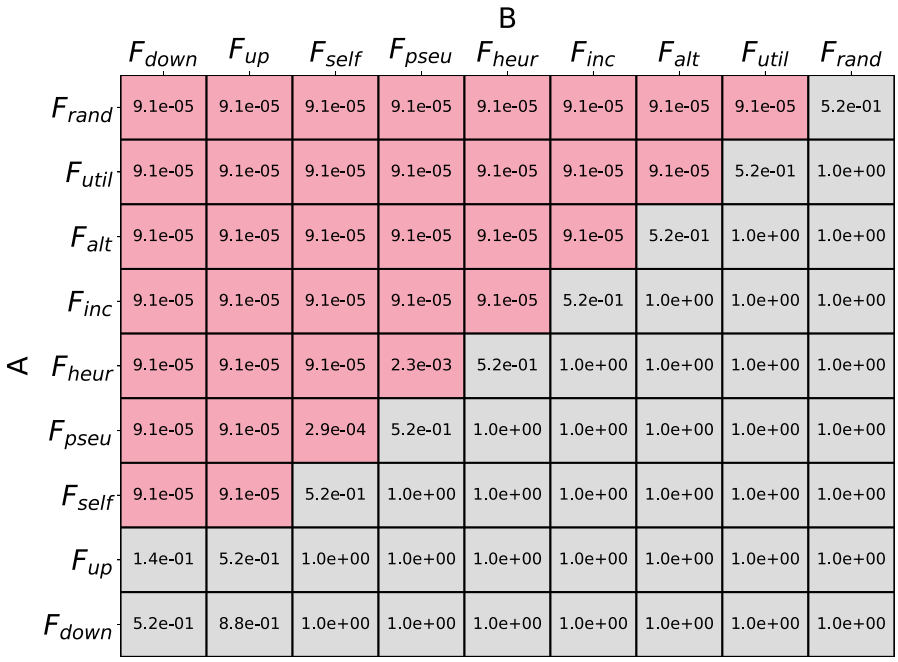

Figure 3.2: Statistical confirmation of the existence of order by importance among factors. Results from systematic Mann-Whitney U tests ($\alpha = 0.05$) comparing the permutation importance of each factor A with every other factor B with H_0 : importance of A = importance of B and H_1 : importance of A $>$ importance of B. Colored cells indicate acceptance of H_1. The results show a clear ordering of factors by importance.

We also desire to assess factor importance in terms of multiple interacting factors, as opposed to analyzing the importance of only singular factors. Figure 3.3 compares the top ten joint contributions to fitness prediction of the random forest by individual factors and joint contributions of factors considered in pairs and triples. F_{rand} far exceeds and other factor or factor interaction in terms of importance. The runners up are $(F_{\text{rand}}, F_{\text{util}})$, $(F_{\text{heur}}, F_{\text{rand}}, F_{\text{util}})$, $(F_{\text{alt}}, F_{\text{rand}})$, $(F_{\text{alt}}, F_{\text{rand}}, F_{\text{inc}})$,

$(F_{\text{rand}}, F_{\text{inc}}, F_{\text{util}})$, and $(F_{\text{alt}}, F_{\text{pseu}}, F_{\text{rand}})$. Note that F_{rand} appears in all of these.

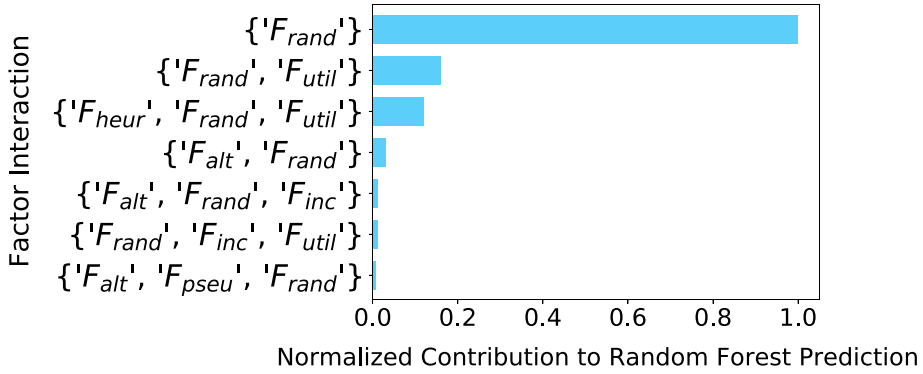

Figure 3.3: Ordered bar chart of the highest normalized joint contribution scores of factors and interactions of three or fewer factors. Above all other factors, pairs, and triplets, F_{rand} alone shows an immensely superior contribution to the random forest's ability to predict model fitness.

Finally, the 14 best-fit candidate models evolved by the genetic program are reported in Table 3.2 along with their mean fitness score. As expected from the factor analysis, F_{rand} is the most frequently appearing factor. Figure 3.4 displays a selection of these top models (rules 1, 3, and 4) compared against a model consisting of just F_{rand} (as a kind of null model) and the original utilitarian model. The EMD-derived models score considerably higher in fitness. We show visualizations of the evolved syntax trees of these models in Figure 3.5.

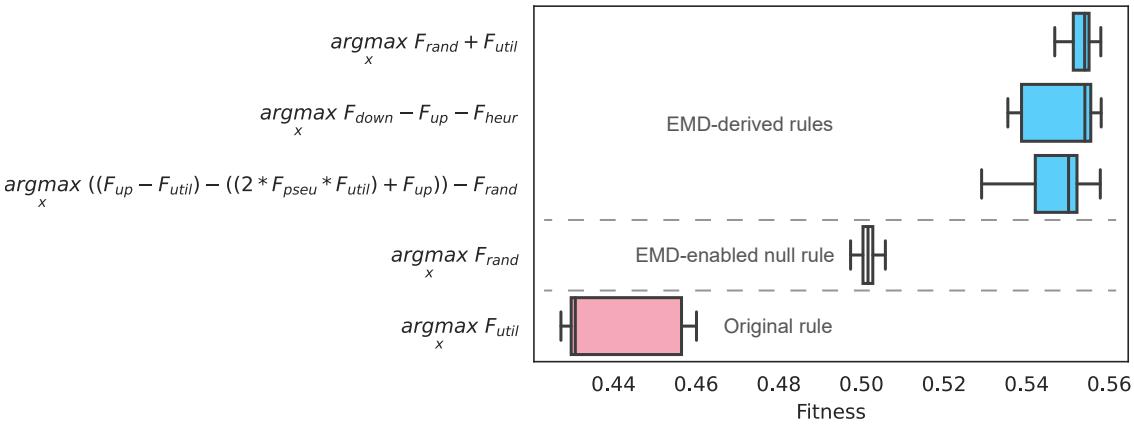

Figure 3.4: Comparisons of 100 samples of three of the top-performing models evolved through EMD compared against a purely random model and the original utilitarian model. Parameters are randomly initialized ± 0.05 about their original optimal values. Models designed through EMD insights are significantly more accurate and robust compared to the original model.

Discussion

In this study, we built upon previous work in the study of irrigation systems social dilemmas. We reanalyzed and augmented a set of agent-based models designed to model behavior in a stylized irrigation dilemma as simulated via a behavioral laboratory experiment. We focused on an individual facet of the model: agents' behavior in contributing to the collective upkeep of irrigation infrastructure. Utilizing evolutionary model discovery, we algorithmically explored a much larger space of possible rule sets for this behavior than was previously possible. We find that, in contrast to the original models, our best-performing rule sets typically have an additional element of stochasticity and favor factors such as other-regarding preferences and perceived relative income.

This idea of baseline behavior augmented with a small amount of additional stochasticity was, in fact, expressed and hypothesized in the models' original introduction (Baggio and Janssen, 2013). For example, rule 1 in Table 3.2 is a simple addition of the baseline utilitarian model with

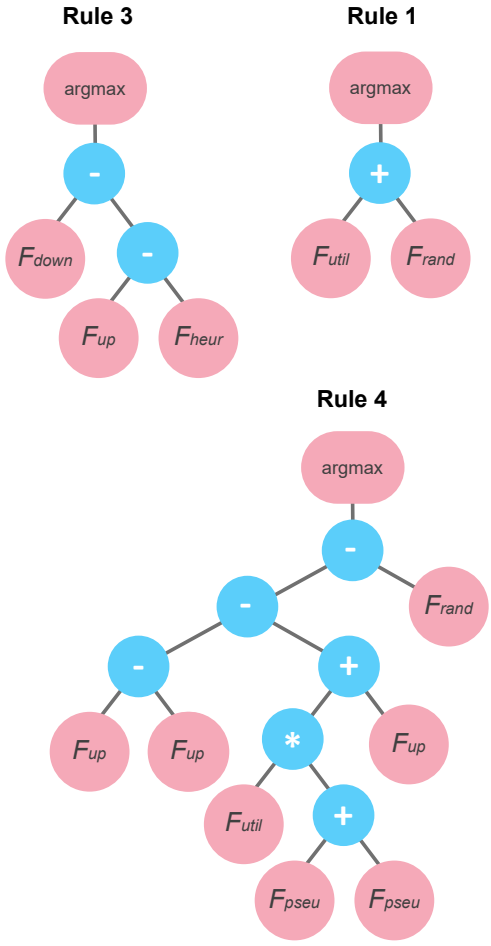

Figure 3.5: Visualizations of the evolved syntax trees for rules 1, 3, and 4.

a random term. Baggio and Jannsen conceptualized this idea as the "wavering hand:" Agents have some idea of how to reason about the system and make decisions, but incomplete understandings and/or uncertainty can cause an amount of stochasticity about a more clearly explained behavior. However, we suspect that the discovery of the robustness of this fact may not have originally occurred as exploring additional rule configuration possibilities was simply too labor-intensive a process prior to modern iGSS methods.

There are additional insights to be had from the models generated and subsequent factor analysis. We note that rule 3 in Table 3.2 represents an interesting outlier in the top performers, as it defies the factor composition which would be expected from the importance analysis. This interaction between the heuristic and both up- and down-stream homophily may belie a particular strategy where upstream farmers are especially motivated to match the investments of downstream farmers. This insight may be used to inform further research on related irrigation dilemma strategies.

Further, we would like to note that, in this current formulation, all agents homogeneously follow the same strategy. This may be a contributing factor to the appearance of disparate rule sets with similar fitness in Table 3.2. Further research may find additional fitness gains through exploring increasingly heterogeneous agents. Additionally, a clear direction forward may also involve allowing the extraction behavior to vary as well as the investment behavior.

In terms of the methodology itself, we would like to note that there is an emerging spectrum which must be balanced in the creation of factors and operators. On one end of the spectrum, primitives take the form of the most basic mathematical and logical building blocks. In theory, any model in existence which is expressible as an equation or computer program could be built using these primitives, and, indeed, Greig and Arranz (2021) succeeded in creating impressively accurate ABMs of physical phenomena through doing so. However; Using primitives like these create what is perhaps the largest possible search space for iGSS algorithms. Immense computational power is required to search it to derive anything but the simplest of models, and different search algorithms will of course produce varying results. On the other end of the spectrum lies constructing primitives and operators which rely heavily upon domain knowledge of the given problem. For example, in our own study, we have ”likelihood of a given investment due to other-regarding-preference-based approximate utility maximization” as a single primitive. Using such primitives can significantly narrow the model search space to only include aspects which the modeller suspects have a high chance of being in the final model. This, of course, has the drawback of the modeller perhaps

narrowing the search space too much and potentially excluding classes of desirable models which would have otherwise been discovered using more basic primitives. Striking the balance between these two extremes seems to be a difficult, albeit promising modelling strategy.

Ultimately, given our results and those of other works (Gunaratne, 2019; Gunaratne et al., 2020; Garibay et al., 2021), we argue that iGSS methods pave a clear path forward for agent-based modelling and represent a natural evolution of the practice. We anticipate that further developments of iGSS will serve to bring about a new era of generative social science models which are not only more accurate and robust, but may aid in uncovering causal mechanisms that otherwise would have remained un-discovered.

Table 3.1: Hypothesized factors contributing to investment behavior

Factor	Name	Description
	Rational factors	
F_{self}	Selfishness	Higher probability for investments closer to 0 tokens
F_{rand}	Random	Uniform random probability
F_{pseu}	Pseudorandom	Pseudorandom "trembling hand" model from Baggio and Janssen (2013); The first investment is randomly chosen, and subsequent investments are the same as the first investment plus or minus a noise term
	Social factors	
F_{alt}	Altruism	Higher probability for investments closer to the maximum token investment
F_{util}	Other-regarding preferences (utilitarian)	Investment behavior of the original base model, introduced in Baggio and Janssen (2013) based on findings from behavioral economics
F_{up}	Upstream homophily	Higher probability for investments more similar to upstream neighbor's last investment
F_{down}	Downstream homophily	Higher probability for investments more simliar to downstream neighbor's last investment
	Emotional factors	
F_{inc}	Relative income	Greater weight for below-average-income agents to invest less and above-average-income agents to invest more.

Table 3.2: The 14 top-scoring rules. Rules are algebraically simplified where applicable. Note that repeat occurences of F_{rand} and F_{pseu} are marked with additional numbered subcripts to highlight the fact that these are non-deterministic functions with values that change each time they are called. Thus, they cannot be algebraically cancelled. For example, $F_{rand} - F_{rand}$ may resolve to 0.5 if the first F_{rand} rolls 0.75 and the second rolls 0.25, so we express this as $F_{rand1} - F_{rand2}$.

#	**Rule**	**Mean fitness**
0	$\underset{x}{\text{argmin}} \quad F_{pseu1} - F_{rand} + F_{util} - F_{pseu2}$	0.5534
1	$\underset{x}{\text{argmax}} \quad F_{rand} + F_{util}$	0.5533
2	$\underset{x}{\text{argmin}} \quad F_{rand} - F_{util}$	0.5531
3	$\underset{x}{\text{argmax}} \quad F_{down} - F_{up} - F_{heur}$	0.5497
4	$\underset{x}{\text{argmin}} \quad ((F_{up} - F_{util}) - ((2 * F_{pseu} * F_{util}) + F_{up})) - F_{rand}$	0.5471
5	$\underset{x}{\text{argmax}} \quad F_{rand1} + (F_{util} * F_{rand2})$	0.5459
6	$\underset{x}{\text{argmin}} \quad F_{pseu} + F_{down} - F_{rand1} - F_{rand2}$	0.5442
7	$\underset{x}{\text{argmin}} \quad (F_{up} * F_{inc}) + F_{rand}$	0.5368
8	$\underset{x}{\text{argmin}} \quad - F_{rand1} - F_{rand2} - F_{pseu} + 2 * F_{inc}$	0.5342
9	$\underset{x}{\text{argmin}} \quad F_{rand1} - F_{rand2} + F_{pseu}$	0.5341
10	$\underset{x}{\text{argmin}} \quad F_{pseu} + F_{rand}$	0.5312
11	$\underset{x}{\text{argmax}} \quad F_{rand} - F_{pseu}$	0.5312
12	$\underset{x}{\text{argmin}} \quad F_{rand} + F_{pseu}$	0.5311
13	$\underset{x}{\text{argmin}} \quad F_{pseu} - F_{rand}$	0.5311

CHAPTER 4: CONCLUSION

In this thesis, we advanced two case studies of artificial social cognition using causal generative models. In the first case study, we generalized a new framework for artificial intent recognition, the Naïve Utility Calculus (NUC), to outperform existing models in observing *stag-hunt*, a simple multi-agent game where agents must infer each others' intent to cooperate to maximize their rewards. In the second case study, we utilized a new paradigm for causal generative modelling known as inverse generative social science (iGSS) to advance models of decision-making in a stylized irrigation system commons dilemma. Using evolutionary model discovery (EMD), an algorithm for iGSS, we found a host of new, algorithmically-deduced models which likewise outperformed the former state-of-the-art models for the problem.

In the stag-hunt case study, NUC displayed a major strength in being the first modelling strategy to outperform the ability of human observers in determining which pairs of agents would decide to cooperate after the first round of observation. While it also displayed some weaknesses, such as a bias in being overconfident of an agent moving in an established direction, the fundamental underpinnings of the model show much potential for continued use in advancing artificial intent recognition.

In the irrigation systems study, we used iGSS methods to study the behavior of farmers in contributing to upkeep of a simulated irrigation canal. This utilized data collected from a series of real-world laboratory experiments wherein participants took on the role of farmers in a stylized irrigation system. Using this data, iGSS-generated models were validated against actual time series describing this behavior, and a host of new models were discovered which fit this data more accurately and robustly than the originally formulated models. The newer models typically incorporated an additional element of stochasticity that was not originally accounted for to the extent

required to achieve such an accurate fit.

These results have clear implications for the advancement of artificial intent recognition and the study of irrigation systems. But, moreover, the work showcases how flexible the methodology of causal generative models and iGSS can be. By virtue of the computational nature of the methodology, each model it puts forth simultaneously advances the scientific domain it is applied to and allows such models to be encoded directly into the cognitive architecture of social AI. Our results show the strength of this multi-front approach to the problem of cooperating to solve our world's largest problems.

We posit that this variety of modeling should continue to be developed as a powerful tool in our collective toolbox to build out more sustainable, equitable, and compassionate societies.

APPENDIX: UCF IRB DETERMINATION

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138, IRB00012110
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

NOT HUMAN RESEARCH DETERMINATION

April 12, 2022

Dear Lux Miranda:

On 4/12/2022, the IRB reviewed the following protocol:

Type of Review:	Initial Study
Title of Study:	Understanding Cooperation and Creating Social AI with Causal Generative Models
Investigator:	Lux Miranda
IRB ID:	STUDY00004198
Funding:	None
Grant ID:	None
Documents Reviewed:	• Miranda_HRP-251.pdf, Category: Faculty Research Approval; • Explanation of analyzed data variables, Category: Other; • MIRANDA_HRP-250.docx, Category: IRB Protocol

The IRB determined that the proposed activity is not research involving human subjects as defined by DHHS and FDA regulations.

IRB review and approval by this organization is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities are research involving human in which the organization is engaged, please submit a new request to the IRB for a determination. You can create a modification by clicking **Create Modification / CR** within the study.

If you have any questions, please contact the UCF IRB at 407-823-2901 or irb@ucf.edu. Please include your project title and IRB number in all correspondence with this office.

Sincerely,

Katie Kilgore
Designated Reviewer

REFERENCES

Anderies, J. M., Janssen, M. A., Bousquet, F., Cardenas, J. C., Castillo, D., Lopez, M. C., Tobias, R., Vollan, B., and Wutich, A. (2011). The challenge of understanding decisions in experimental studies of common pool resource governance. *Ecological Economics*, 70(9):1571–1579.

Anderies, J. M., Janssen, M. A., Lee, A., and Wasserman, H. (2013). Environmental variability and collective action: Experimental insights from an irrigation game. *Ecological Economics*, 93:166–176.

Baggio, J. A. and Janssen, M. A. (2013). Comparing agent-based models on experimental data of irrigation games. In *2013 Winter Simulations Conference (WSC)*, pages 1742–1753, Washington, DC, USA. IEEE.

Baggio, J. A., Rollins, N. D., Pérez, I., and Janssen, M. A. (2015). Irrigation experiments in the lab: Trust, environmental variability, and collective action. *Ecology and Society*, 20(4):12.

Barnes, M., Chen, J., Schaefer, K. E., Kelley, T., Giammanco, C., and Hill, S. (2017). Five Requisites for Human-Agent Decision Sharing in Military Environments. In Savage-Knepshield, P. and Chen, J., editors, *Advances in Human Factors in Robots and Unmanned Systems*, pages 39–48, Cham. Springer International Publishing.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, 4(1):15–31.

Cifdaloz, O., Regmi, A., Anderies, J. M., and Rodriguez, A. A. (2010). Robustness, vulnerability, and adaptive capacity in small-scale social-ecological systems: The pumpa irrigation system in nepal. *Ecology and Society*, 15(3).

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, 31(1):169–193.

Demiris, Y. (2007). Prediction of intent in robotics and multi-agent systems. *Cognitive Processing*, 8(3):151–158.

Dietz, T., Ostrom, E., and Stern, P. C. (2003). Struggle to Govern the Commons. *Science*, 302(5652):1907–1912.

Elsawah, S., Filatova, T., Jakeman, A. J., Kettner, A. J., Zellner, M. L., Athanasiadis, I. N., Hamilton, S. H., Axtell, R. L., Brown, D. G., Gilligan, J. M., Janssen, M. A., Robinson, D. T., Rozenberg, J., Ullah, I. I. T., and Lade, S. J. (2020). Eight grand challenges in socio-environmental systems modeling. *Socio-Environmental Systems Modelling*, 2:16226.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, 4(5):41–60.

Epstein, J. M. (2014). *Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science*. Princeton University Press. Publication Title: Agent_Zero.

Falkenhainer, B., Forbus, K. D., and Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1):1–63.

Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., and Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors*, 52(2):203–224.

Fiore, S. M. and Wiltshire, T. J. (2016). Technology as teammate: Examining the role of external cognition in support of team cognitive processes. *Frontiers in Psychology*, 7(OCT).

Forbus, K. D., Ferguson, R. W., Lovett, A., and Gentner, D. (2017). Extending SME to Handle Large-Scale Cognitive Modeling. *Cognitive Science*, 41(5):1152–1201.

Freeman, J., Baggio, J. A., and Coyle, T. R. (2020). Social and general intelligence improves collective action in a common pool resource system. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14):7712–7718.

Garibay, I., Epstein, J. M., and Rand, W. (2021). Inverse generative social science workshop. `https://www.igss-workshop.org`.

Garibay, I., Oghaz, T. A., Yousefi, N., Mutlu, E. C., Schiappa, M., Scheinert, S., Anagnostopoulos, G. C., Bouwens, C., Fiore, S. M., Mantzaris, A., et al. (2020). Deep agent: Studying the dynamics of information spread and evolution in social networks. *arXiv preprint arXiv:2003.11611*.

Greig, R. and Arranz, J. (2021). Generating agent based models from scratch with genetic pro-

gramming. In *ALIFE 2021: The 2021 Conference on Artificial Life*. MIT Press.

Gunaratne, C. (2019). *Evolutionary Model Discovery: Automating Causal Inference for Generative Models of Human Social Behavior*. PhD thesis, University of Central Florida.

Gunaratne, C., Baral, N., Rand, W., Garibay, I., Jayalath, C., and Senevirathna, C. (2020). The effects of information overload on online conversation dynamics. *Computational and Mathematical Organization Theory*, 26(2):255–276.

Gunaratne, C. and Garibay, I. (2017). Alternate social theory discovery using genetic programming: towards better understanding the artificial anasazi. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 115–122. ACM.

Gunaratne, C. and Garibay, I. (2020). Evolutionary model discovery of causal factors behind the socio-agricultural behavior of the Ancestral Pueblo. *PLOS ONE*, 15(12):e0239922. Publisher: Public Library of Science.

Gunaratne, C., Rand, W., and Garibay, I. (2021). Inferring mechanisms of response prioritization on social media under information overload. *Scientific reports*, 11(1):1–12.

Gutiérrez, N. L., Hilborn, R., and Defeo, O. (2011). Leadership, social capital and incentives promote successful fisheries. *Nature*, 470(7334):386–389.

Janssen, M. A., Anderies, J. M., Pérez, I., and Yu, D. J. (2015). The effect of information in a behavioral irrigation experiment. *Water Resources and Economics*, 12:14–26.

Janssen, M. A. and Baggio, J. A. (2017). Using agent-based models to compare behavioral theories on experimental data: Application for irrigation games. *Journal of Environmental Psychology*, 52:194–203.

Janssen, M. A., Bousquet, F., Cardenas, J.-C., Castillo, D., and Worrapimphong, K. (2012). Field experiments on irrigation dilemmas. *Agricultural Systems*, 109:65–75.

Janssen, M. A., Gharavi, L., and Yichao, M. (2020). Keeping Up Shared Infrastructure on a Port of Mars: An Experimental Study. *International Journal of the Commons*, 14(1):404.

Jara-ettinger, J., Gweon, H., Schulz, L. E., and Tenenbaum, J. B. (2016). The Naïve Utility Cal-

culus : Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, 20(8):589–604.

Jara-ettinger, J., Gweon, H., Tenenbaum, J. B., and Schulz, L. E. (2015). Children ' s understanding of the costs and rewards underlying rational action. *Cognition*, 140:14–23.

Jara-Ettinger, J., Schulz, L. E., and Tenenbaum, J. B. (2020). The Naïve Utility Calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123.

Johnson, M., Hofmann, K., Hutton, T., and Bignell, D. (2016). The malmo platform for artificial intelligence experimentation. *IJCAI International Joint Conference on Artificial Intelligence*, 2016-Janua:4246–4247.

Michelozzi, P. and De'Donato, F. (2021). Ipcc sixth assessment report: stopping climate change to save our planet.

Miranda, L. and Freeman, J. (2020). The two types of society: Computationally revealing recurrent social formations and their evolutionary trajectories. *PLOS ONE*, 15(5):e0232609. Publisher: Public Library of Science.

Miranda, L. and Ozmen Garibay, O. (2021). Multi-agent Naive Utility Calculus: Intent Recognition in the Stag-Hunt Game. In Thomson, R., Hussain, M. N., Dancy, C., and Pyke, A., editors, *Social, Cultural, and Behavioral Modeling*, Lecture Notes in Computer Science, pages 331–340, Cham. Springer International Publishing.

Orr, M. G., Lebiere, C., Stocco, A., Pirolli, P., Pires, B., and Kennedy, W. G. (2018). Multi-scale Resolution of Cognitive Architectures: A Paradigm for Simulating Minds and Society. In Thomson, R., Dancy, C., Hyder, A., and Bisgin, H., editors, *Social, Cultural, and Behavioral Modeling*, pages 3–15, Cham. Springer International Publishing.

Ortman, S. G. (2018). Cultural Genotypes and Social Complexity. In Sabloff, J. A. and Sabloff, P. L. W., editors, *The Emergence of Premodern States*, pages 221–268. SFI Press.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

Ostrom, E., Burger, J., Field, C. B., Norgaard, R. B., and Policansky, D. (1999). Revisiting the commons: local lessons, global challenges. *science*, 284(5412):278–282.

Qi, S. and Zhu, S. (2018). Intent-aware multi-agent reinforcement learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7533–7540.

Rabkina, I. (2020). *Analogical Theory of Mind: Computational Model and Applications*. PhD thesis, Northwestern University.

Rabkina, I. and Forbus, K. D. (2019). Analogical Reasoning for Intent Recognition and Action Prediction in Multi-Agent Systems. In *Proceedings of the 7th Annual Conference on Advances in Cognitive Systems*.

Rajabi, m., Gunaratne, C., Mantzaris, A. V., and Garibay, I. (2020). On countering disinformation with caution: Effective inoculation strategies and others that backfire into community hyperpolarization. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 130–139. Springer.

Schelling, T. C. (1969). Models of segregation. *The American economic review*, 59(2):488–493.

Schlüter, M., Baeza, A., Dressler, G., Frank, K., Groeneveld, J., Jager, W., Janssen, M. A., McAllister, R. R., Müller, B., Orach, K., Schwarz, N., and Wijermans, N. (2017). A framework for mapping and comparing behavioural theories in models of social-ecological systems. *Ecological Economics*, 131:21–35.

Scholkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634.

Shum, M., Kleiman-Weiner, M., Littman, M. L., and Tenenbaum, J. B. (2019). Theory of minds: Understanding behavior in groups through inverse planning. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 6163–6170.

Skyrms, B. (2003). The stag hunt and the evolution of social structure. *The Stag Hunt and the Evolution of Social Structure*, pages 1–149.

Sukthankar, G., Geib, C., Bui, H. H., Pynadath, D., and Goldman, R. P. (2014). *Plan, activity, and intent recognition: Theory and practice*. Newnes.

Taleb, N. N. (2012). *Antifragile: Things that gain from disorder*, volume 3. Random House.

Turchin, P. (2016). *Ultrasociety: How 10,000 years of war made humans the greatest cooperators on earth*. Beresta Books Chaplin, CT.

Vu, T. M., Probst, C., Epstein, J. M., Brennan, A., Strong, M., and Purshouse, R. C. (2019). Toward inverse generative social science using multi-objective genetic programming. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1356–1363, Prague Czech Republic. ACM.

Winkle, K., Senft, E., and Lemaignan, S. (2021). LEADOR: A Method for End-To-End Participatory Design of Autonomous Social Robots. *Frontiers in Robotics and AI*, 8:704119.