



Does big data serve policy? Not without context. An experiment with in silico social science

Chris Graziul¹ · Alexander Belikov¹ · Ishanu Chattopadhyay¹ · Ziwen Chen¹ · Hongbo Fang² · Anuraag Girdhar¹, et al. *[full author details at the end of the article]*

Accepted: 24 May 2022 / Published online: 30 November 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

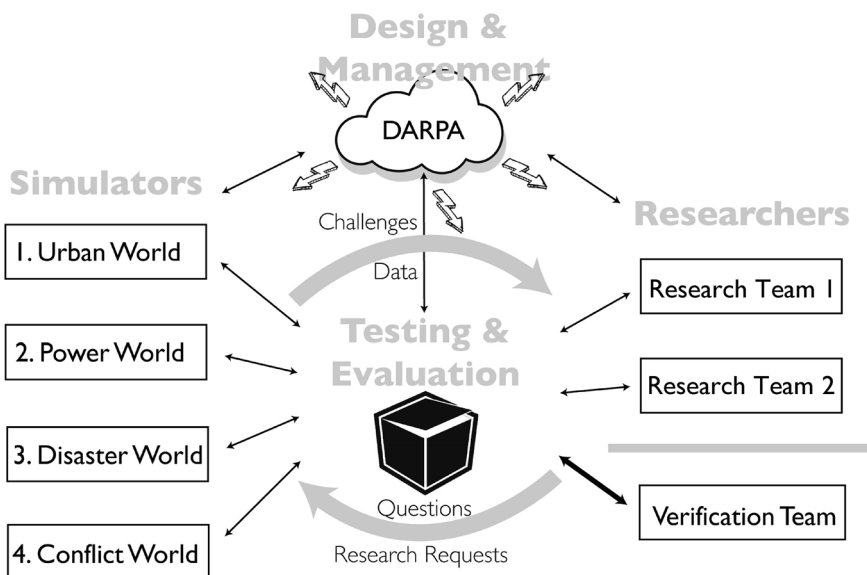
The DARPA Ground Truth project sought to evaluate social science by constructing four varied simulated social worlds with hidden causality and unleashed teams of scientists to collect data, discover their causal structure, predict their future, and prescribe policies to create desired outcomes. This large-scale, long-term experiment of in silico social science, about which the ground truth of simulated worlds was known, but not by us, reveals the limits of contemporary quantitative social science methodology. First, problem solving without a shared ontology—in which many world characteristics remain existentially uncertain—poses strong limits to quantitative analysis even when scientists share a common task, and suggests how they could become insurmountable without it. Second, data labels biased the associations our analysts made and assumptions they employed, often away from the simulated causal processes those labels signified, suggesting limits on the degree to which analytic concepts developed in one domain may port to others. Third, the current standard for computational social science publication is a demonstration of novel causes, but this limits the relevance of models to solve problems and propose policies that benefit from the simpler and less surprising answers associated with most important causes, or the combination of all causes. Fourth, most singular quantitative methods applied on their own did not help to solve most analytical challenges, and we explored a range of established and emerging methods, including probabilistic programming, deep neural networks, systems of predictive probabilistic finite state machines, and more to achieve plausible solutions. However, despite these limitations common to the current practice of computational social science, we find on the positive side that even imperfect knowledge can be sufficient to identify robust prediction if a more pluralistic approach is applied. Applying competing approaches by distinct subteams, including at one point the vast TopCoder.com global community of problem solvers, enabled discovery of many aspects of the relevant structure underlying worlds that singular methods could not. Together, these lessons suggest how different a policy-oriented computational social science would be than the

computational social science we have inherited. Computational social science that serves policy would need to endure more failure, sustain more diversity, maintain more uncertainty, and allow for more complexity than current institutions support.

Keywords Computational social science · Simulated societies · Policy · Quantitative social science · Machine learning · Deep learning · Simulation

1 Let the expedition begin

The DARPA Ground Truth (GT) project provided an abundance of novel and unexpected opportunities for participants to select and implement similarly novel methodological approaches. Its stated purpose was to test the viability of using simulated social systems to conduct productive social scientific research. To advance this goal, hundreds of researchers and key support personnel participated in the GT project in various capacities. The project was highly structured in how researchers were organized into groups, how these groups did (or did not) communicate, what information was (or was not) provided to each group, and what problems each group had to solve. Otherwise, participants were free to choose how to achieve their goals. See Fig. 1 for an overview.



Note: Participating Simulator and Researcher teams were often comprised of multiple formal or informal sub-groups but, for brevity, are not depicted here.

Fig. 1 DARPA ground truth project collaborative research strategy

1.1 Four worlds

The formal structure of the GT project, in terms of research teams and their roles, involved the creation of one simulated social system by each of four *simulation* teams (TA1A-D), which our team affectionately titled “Urban World”, “Power World”, “Disaster World” and “Conflict World”. These four “worlds” were then explored and studied by two different *research* teams (TA2A-B). The research teams who explored and studied these worlds had no knowledge about the simulator teams who created them nor, for the first year, each other. Our team—one of the research teams—employed a pluralistic approach that sought to engage many possible methods and models. The other research team focused on a mixed methods design, applying and extending advances in causal analysis, conducting sociologically informed modeling, and using agent-based modeling to replicate phenomena of interest. A seventh testing and evaluation team (T&E) provided guidance to those building the simulations about the substantive features of each world, determined what data could be shared by simulation teams with research teams, and evaluated the performance of both simulators and researchers. Within this structure, T&E was understood to represent the research interests and intended goals of DARPA as the organization funding this research. Finally, an eighth team from the Pacific Northwest National Laboratory was assigned to replicate results submitted by the research teams during the final phase of the GT project.

As can be read in other papers of this issue of *CMOT*, the four worlds were varied in substance and mechanism. “Urban world” was created by a simulation team of quantitative geographers from George Washington University on the street grid of a modern city and was powered by an agent-based model that balanced personal preferences and exigencies of locations, such as work, recreation sites, restaurants and home. Individuals had money, made friends, and could eventually contract disease, and research tasks surrounded the creation of more (or less!) income equity, more social connections, and lower morbidity. “Power world” was created by an industry-led engineering team at Raytheon in collaboration with social scientists based on principles of social groups and complex collective behavior. Also powered by an agent-based model, Power World involved regional elections and policies, group competition, and group-acquired income. Individuals exhibited levels of happiness, and research tasks involved the creation of policies that increased happiness while achieving particular election outcomes and other group success criteria. “Disaster World” was created by machine learning and artificial intelligence researchers from USC’s Institute for Creative Technologies who developed an agent-based model of human behavior in response to risks associated with a hurricane, which was driven probabilistically with partially observable Markov decision processes (POMDPs). Disaster World inhabitants maximized their “reward” (i.e. personal health status, the health status of family members, etc.) by reacting to perceived risks and realized outcomes associated with different courses of action vis-a-vis hurricane impact. “Conflict World” was created by a team from Wright State research institute in

partnership with experienced intelligence analysts to generate an agent-based model that reflected a state in crisis, with civil conflict, violent insurgency, food shortages and popular unrest. Their model represented the synthesis of multiple types of perspectives about the state of the world, possible courses of action, and the role of path dependency. Its defining feature was a simulation where actors' preferences and choices responded to a dynamic shared environment.

The GT project timeline consisted of 30 months of activities organized around three project phases that each consisted of three tasks. These tasks were formally labeled Explain, Predict, and Prescribe, corresponding to the type of questions or problems research teams were expected to answer or solve. While these tasks were uniform in character across phases, it was understood that simulation teams would make their worlds more complex with each phase.

Aside from the common use of agent-based models to generate data, the content and structure of the GT project's four worlds differed vastly and in fundamental ways. For example, agents in Power World sought to maximize their utility, simulating the familiar *Homo economicus* located in an unfamiliar landscape of opportunity (Granovetter 1985). Conversely, agents in Conflict World engaged in stigmergic behavior (Padgett and Powell 2012), actively shaping and responding to their simulated environment in a manner characteristic of *Homo sociologus* (Dahrendorf 1973), an agent driven by personally acquired or cultivated values in a landscape of dynamic, shifting possibilities. This range in simulated characteristics of social systems, from market-based utility maximization to culture-based co-constitution of agents and their environment, was not known to research teams during the GT project but reflects extreme variation in assumptions made about human behavior within social systems. This variety affected not only the properties of agents, such as their motivations, but other properties of the simulations, such as the kinds of social processes that existed in each world.

Within this framework, research teams sought to validate existing quantitative tools and test novel methodologies to conduct social science on these simulations, DARPA's stated goal of the GT project. These efforts spanned a range of approaches encompassing traditional social science methods, bespoke and novel methods reliant on sophisticated computational strategies, and bleeding edge neural models from computer science, which we discuss in the final section of the manuscript. Our team also tested the efficacy of crowdsourcing for completing tasks in Phases 1 and 2. This work sought to realize the stated goals of the GT project through any means necessary.

It also seemed plausible to our team the GT project might serve other valuable purposes for DARPA, such as the creation of better wargames or the evaluation of standard social science methodologies' ability to identify a novel process not previously found in the real world but built into the simulation. The ambiguity of potential uses for project outputs created a unique situation where decisions by T&E to restrict or allow the flow of certain kinds of information between simulation and research teams (a) had definitive impacts on how and what research could be conducted and, due to these impacts, (b) inspired discussions about the metagame that may be structuring these decisions. This feature of the GT project is notable as most

large-scale social science research involves proposing a specific, highly detailed research plan with clear aims that align with the funding organization's goals. While we were provided specific tasks, our performance was evaluated using known metrics, and project goals were documented/discussed explicitly, we perceived ambiguity in project goals (i.e., the potential existence of an undocumented and undiscussed metagame) which led to real ambiguity about what we believed to constitute productive versus unproductive effort within strict project timelines. The combination of strict timelines, restricted information flows, and the assumed known unknown of the GT project's potential "true" goal(s) presented a novel problem-solving environment that negatively affected our team yet inspired substantial reflection about the basic nature of social scientific research tasks such as survey data collection and ethnography.

Along with strict timelines, teams participating in the GT project experienced other constraints which shaped lines of inquiry and thus generation of knowledge about each world. First, simulation teams created initial data for analysis by research teams. In Phase 1 and 2, these initial data were incomplete, intentionally excluding the vast majority of data generated for each world since data collection was a major part of research teams' tasks. To complete data collection tasks, research teams posed well-constructed questions and requests for data. Within this framework, questions were used as a way to learn about the existence (or properties) of objects in a simulated world, while requests for data, termed "research requests," were used to simulate the implementation of social scientific research methods to collect data for analysis. Submitted questions and research requests had to adhere to strict guidelines regarding their structure, content, and framing. In Phase 3 this constraint was removed, with simulation teams providing all generated data to research teams, precipitating the inverse problem of data management (i.e., analysis of hundreds of gigabytes of high frequency simulation data). In both cases, the novelty of the problem-solving environment again inspired self-reflection about the idiosyncratic nature of social scientific inquiry as it has evolved in the real world, particularly its tendency to expect certain inputs for analysis. Specifically, we learned tacit disciplinary knowledge about human behavior led to the expectation that (a) certain features existed in the simulation because they were so "fundamental" to understanding agent behavior or, conversely, (b) certain features of human experience were so basic (e.g., feeling hungry) they could be ignored, both of which revealed how difficult it was for us to decontextualize simulated human behavior.

A second constraint faced by simulation and research teams was the fact that the research request system was managed by the T&E team, a third-party mediating how teams interacted. Research requests (RRs) generated by research teams were expected to reflect actual methods of research used within the social sciences, such as fielding a survey asking a series of questions about agents' age, current status, and preferences. In this context, each world had a temporal component similar to sampling frequency, which sometimes meant that well-formed RRs inadvertently led to the generation of massive amounts of data. For example, Urban World's agents were observed in 5-minute increments over a period of multiple simulated years. Not knowing the temporal scale of (sub)-mechanisms that governed these agents' behavior meant it was often safest to generate RRs with maximum sampling

frequency. Related, when simulation teams received RRs they had to both (a) interpret the intended meaning of the proposed data request and (b) conform to limits placed on the kind and amount of data they could provide according to the T&E team. This system enabled “blinding” of simulation/research teams by making T&E the intermediary between both. This constraint introduced yet another novel feature of the problem-solving environment in the form of “blinded” data collection which required self-reflection about the biases we held about the data collection process. We found through trial and error that biases about which features to try to measure (and further study) could, again, turn assets into liabilities, expertise into folly, if we failed to accept certain seemingly natural lines of inquiry were unproductive.

Within these constraints, the GT project consisted of three distinct tasks during each phase. Each task focused on producing a particular kind of knowledge:

The “Explain” task required research teams to develop as full an understanding of the causal relationships within each simulated world (i.e., a directed causal graph) as possible. This task required identification of nodes consisting of features within the simulation, such as agent age and current location, and edges consisting of causal relationships between features, such as the tendency for older agents to visit particular sites. Fulfilling this task required discovering all features within each world and performing an exhaustive search for all causal relationships between these features.

The “Predict” task required research teams to make an out of sample prediction of agent behavior—individually, in aggregate, or as an observable property of the simulated world—based on data provided by or solicited from simulation teams. This task always followed the Explain task and so, while identification of additional features and causal relationships sometimes continued, efforts often focused on converting a working knowledge of each simulated world into a model of each social system. The primary requirement of this model was fidelity to the ground truth of each simulated world. Fulfilling this task required discovering methods that could accurately reproduce observed outcomes, such as agent behavior, within each world. Nevertheless, some prediction tasks were counterfactual—if *X* happened tomorrow, what would happen next year?—demanding a rich understanding of causal relationships that were supposed to be uncovered during the “Explain” task.

The “Prescribe” task required research teams to optimize a set of possible interventions for each world to maximize or minimize a specific property of the world (e.g., average number of friendships, total number of casualties, etc.). As in the Predict task, team performance was evaluated via out of sample observations: simulation teams implemented the interventions provided by research teams then observed their effects, recording relevant metrics for a period of time following the time series observable by research teams. Fulfilling this task required developing a counterfactual understanding for each simulated world, specifically (a) submission of RRs to provide data on the holistic effect of interventions on various aspects of the social system and (b) development of an optimization strategy. The latter logically focused on the features and causal relationships that most strongly impacted the evaluation metrics before considering secondary and tertiary aspects of the world that perturbed these metrics.

Here we report how the structure of the GT project and the nature of its activities contribute to our understanding of knowledge production about social systems via

our experiences studying the simulated worlds generated by simulation teams. This is from the perspective of our research team, initially branded as Social Machine Intelligence for Novel Discovery (or Social MIND) to reflect our commitment to explore the connection between emerging artificial intelligence techniques and familiar social science goals. Many of our insights are best characterized in terms of ontological, epistemological, and methodological issues raised and accompanying lessons learned. We believe these issues are relevant to knowledge production in the social and behavioral sciences broadly, especially within collective problem solving settings where rapid development of effective public policy is the primary goal, such as during the COVID-19 pandemic.

1.2 Ontology of in silico worlds

The type and extent of ontological issues associated with the GT project were unexpected. Some issues, such as discovery of features and causal mechanisms, stemmed directly from the structure and stated goal of the project. We briefly discuss these expected challenges but focus discussion on broader questions with substantial implications for collective problem solving. These questions arose from parallels between the GT research setting and real world problems, like political decision making about complex social issues or environmental challenges. We generally conclude that ontological questions about the nature of social systems can pose significant risks to the ability of disparate groups to arrive at a consensus about how to solve complex problems, such as racial disparities in public health and appropriate local responses to a global pandemic. We organize these questions according to the type of “game” being played when we study social systems, the ambiguity around what constitutes a social system, and the process of data collection given limited resources.

1.3 Expeditions into the unknown: a game within a game?

Limited initial knowledge about simulation teams’ worlds presented a unique opportunity to test the efficacy of different problem-solving strategies. One novel aspect of this extraordinary ambiguity was the fact that we had no sense of the type of “game” (i.e., rules used to build simulations, goals of each simulation, etc.) being played by simulation teams nor the potential metagame being played by DARPA, though the assumption of an unknown metagame was unique to our team alone. Thus, there was little reason to believe any one disciplinary perspective or methodological skill set would be better suited to achieving GT project tasks than any other. To this end, we tested the efficacy of approaching each task as a lone explorer, as two explorers, and as a team or community expedition into the four worlds.

Multiple worlds were studied as single explorer expeditions where we tested if the most effective way to complete a task was for one individual (or one methodological approach) to direct the bulk of our activities. In our team, each world had either one or two primary “explorers” with additional individuals providing support. A single explorer approach meant the use of a novel methodological strategy whose

execution was led by a single individual, sometimes supported by others acting at their direction. For example, for Urban World we used Granger causal networks to complete tasks in Phase 1 and Phase 2 (Chattopadhyay 2014; Li et al. 2019). This novel method was content free in terms of being broadly applicable to the type of data associated with Urban World regardless of what these data represented. We found that this flexibility sometimes helped ensure certain tasks could be completed, but that it relied on a small handful of experts to produce results which were not always interpretable within our broader understanding of the world. Single explorers were good at finding a way to play the game using their own rules but the idiosyncratic nature of their efforts could limit the generalizability of their modeling strategy even among those with domain knowledge.

Another approach to GT project tasks was to treat these tasks as a two explorer cooperative expedition. A two explorer approach meant that two individuals cooperatively explored and tested different methods for completing tasks. Cooperative exploration and problem solving allowed us to leverage both individuals' perspectives and skill sets to iteratively uncover effective solutions to the problems faced in each phase. We expected that pairing researchers with differing skill sets, differing levels of experiences, or both could be effective for completing GT tasks (Hall et al. 2018).

Overall, work on Urban and Disaster Worlds supported this hypothesis about the nature of the GT project's challenges and the best way to handle inherent uncertainties about simulated worlds. In each case, insights from both explorers contributed significantly to completing tasks depending on the underlying ground truth of simulated relationships. For example, in Urban World one player was well-versed in GIS and spatial analysis, which proved more salient to performing certain tasks, while the second player was versed in social theory and network analysis, which helped provide initial directions for analysis. We elaborate below how this role-taking worked in practice. We observed that even when empirically grounded, theory-based analysis proved uninformative it still allowed us to rule out common theories of social behavior or common modes of social scientific analysis (e.g., the use of exponential random graph models to study friendship networks between agents). Our experience with the GT project suggests that two heads are, indeed, better than one when presented with a complex research task whose parameters are not fully known—which reflects the vast majority of research tasks (Tambe et al. 1999; Denzinger 1995; Xyrichis and Ream 2008).

The final approach we tested for effectively completing GT tasks was to assume each task represented a team or community expedition. In this context, a team expedition involved many individuals working to independently complete a task using their personal understanding of the task and the most suitable methods for completing the task. The “team” aspect of this approach is embodied by shared striving toward a common destination given identical information and constraints. Team expeditions were operationalized through use of crowdsourcing competitions to solicit solutions to the challenges associated with GT tasks. For Phase 2, we worked with TopCoder.com, which maintains a vast community of “solvers” who engage with data science, programming, and other scientific challenges to win prizes for the top solutions. TopCoder represents a large global network

of technologists, boasting approximately 1.5 million community members. Up to 450,000 members of this community are data scientists with backgrounds in computer science, physics, mathematics, and engineering. In Phase 2 of the project, we involved TopCoder in a four tournament Grand Challenge through which we solicited community assistance to: (1) suggest possible social entities, forces, and causal relations in each of the four worlds to stimulate our requests for data, as well as (2) accomplish the explain, predict, and prescribe tasks alongside our teams. We scheduled completion of these challenges a week before each of our solutions were due, with our independent or cooperative explorers intensively engaging with their results to improve our final submissions. Across all three competitions we had 686 registered participants, from 15 countries; 64% from India, 7% from Russia, 5% from Indonesia, 4% from the United States and less than 2% were from Kenya, Iran, China, Philippines, South Africa, Italy, Canada, Romania, Brazil, Mexico and Egypt.

For the first of our four challenges, we received a barrage of suggestions that broadened our scope of questions and research requests to simulation teams. These included using more of the research methods available to us, such as experimentation, social media analyses or use of government records, and the proposition of out-of-the box causal possibilities (e.g. boredom; importance of cultural diversity, public safety, presence of public transit; innate intelligence, actively blogging, having a criminal background, participating in ongoing criminal behavior, experience of bullying, alcohol/drug use, number of languages spoken, fatigue, emotional state, materialism, global warming, pollution, drug addiction, volcanic eruption, etc.).

For the remaining challenges, TopCoder solvers had a more difficult time, at least in part because they had to work with an evolving dataset in a compressed time frame without the benefit of direct feedback on the success of their strategies. We found this problem solving context made the creation of a leaderboard, and the accelerating competition it engenders at the end of a competition, infeasible. This finding illustrates how the features of such contexts shape how effectively crowdsourcing can be leveraged as a form of collective problem solving. Moreover, the tasks were complex and involved many continuously evolving forms of data to produce many different required predictions and prescriptions. We observed it was difficult for solvers to put together a complete solution despite efforts to share code that integrated and cleaned relevant data. In other cases, community solutions misunderstood the data or questions, or added nothing new to own analyses. Other times, community members proposed methods like Bayesian Networks for causal discovery (Cheng et al. 2002) or categorical boosted forests, implementing these methods via working python code that successfully uncovered world dynamics. We then extended these with new data and applied them to those same worlds and others.

Consider Urban World's Phase 2 Prescribe task as an illustration of how these different problem solving strategies worked in practice to help triangulate the best possible submission. The goal of this task was to select a subset of 200 agents in the world who will, over a 30 day period and in isolation (i.e. all other agents in the world are removed), collectively exhibit higher average daily friendship network degree over the final week of this period than any other subset of 200 agents subject to the same conditions. As part of this task, we were provided a sample evaluation

Table 1 Comparison of evaluation metrics in urban world phase 2 prescribe

Source	Average degree (08/08–08/14)	Change
Initial data package	5.05	–
Mock test #1	6.15	+ 22%
Mock test #2	5.87	+ 16%
Mock test #3	6.36	+ 26%
Mock test #4	15.04	+ 298%
Final submission	14.10	+ 279%

metric (i.e. 5.05) in the initial data package representing the outcome of drawing a uniform random sample of 200 agents. We were then allowed to submit up to four mock test sets of 200 agents where the simulation team would then provide the corresponding value of the evaluation metric based on each test set of agents. Table 1 summarizes the results of these mock tests and relative change in the evaluation metric versus the sampling strategy and metric included in the initial data package.

Each of the four mock tests in Urban World's Phase 2 Prescribe task represented a systematic approach to the selection of 200 agents. The first mock test drew a stratified random sample of agents designed to reflect the age, education, and income distributions of all agents, under the belief that these agent-based characteristics, including homophily (McPherson et al. 2001), might systematically affect friendship networks. The second mock test used a matched sampling approach designed to maximize insight into how other features of the world affected friendship networks, such as the role of location. The third mock test focused on the effect of geographic propinquity on friendship networks (e.g. Holzhauer et al. 2013).

The fourth and final mock test represented an entirely new perspective on the selection of agents. The first and second test sets were drawn based on input from a team member trained in social theory and social network analysis, the third test set was developed between this team member and another team member trained in GIS methods, but the fourth set was generated based on intuitions about spatial processes as understood by the GIS-trained team member. The fourth set was derived by applying a clustering algorithm to the location of agents' homes weighted by the "popularity" of each home (i.e. how many friendship ties were linked to a home through its residents). The substantial improvement in evaluation metric over the other three mock tests was exciting and moderately unexpected. However, the overall utility of this clustering approach became apparent when we learned that one of the TopCoder submissions used clustering in a similar yet more intricate manner. This work inspired us to pursue essentially the same strategy used in fourth mock test for our final submission, tweaking our approach based on input from the TopCoder submission that also used clustering. Ultimately, this fine tuning led to a slightly lower final evaluation metric. However, without the TopCoder submission we would have felt more uncertain about the exclusive use of clustering and might have made more

substantial changes to our strategy that would have resulted in a much lower evaluation metric. In this case, the two explorers experimented with strategies based on analytic perspectives cultivated in two distinct fields of study, but the value of one explorer's perspective (i.e. GIS-based clustering) became unambiguous via use of crowdsourcing to externally validate its efficacy.

1.4 Simulated social systems still require real definitions

Ontological questions about what constituted a social system presented both the most difficult challenges to accomplishing GT tasks as well as, arguably, the most insight into social science as a practice. The basic ontological questions that plagued both simulation and research teams was “What exists in the simulated social system?” and “How do objects in the simulated social system interact with each other?” when it was made clear to all involved that these simulations were *not* meant to be realistic representations of social behavior; the GT project recreated the basic features of “first contact” with an unknown civilization (and, at times, used language to suggest that research teams should assume they are studying an alien planet). The setting, then, was intended to be functionally similar to a real world social system where individuals' roles, motivations, and behaviors are not known to the observer.

Determining what objects exist and how these objects interact was the most fundamental task in the GT project; without this information it is impossible to explain/predict behavior or prescribe interventions. Because GT worlds were simulated social systems whose features were artificially generated and intentionally *not* reflective of real social systems, and because research teams had to communicate with simulation teams to request additional data, who in turn had to interpret these requests in order to provide the requested data, the interpretation of these requests proved to be fundamental for making sense of each world. Basic ontological questions about what objects exist and how these objects interact were thus complicated by each team's use of words in naming data attributes. For example, the team behind Urban World explicitly renamed variables in Phase 2 to prevent observed cases in Phase 1 where research teams had inferred erroneous information based on variable names (e.g., assuming that “has Child” meant the existence of family units when, in fact, “has Child” was only meant to indicate that agents had increased expenses if they had a child).

Another ontological question raised by GT project tasks was the overall purpose of each world or, more properly, the win condition(s) of each game embedded within these four worlds. For example, Conflict World presented unique challenges to our team because we assumed that all simulated agents followed a basic utility maximization principle. While more or less true in other worlds, the stigmergic basis of Conflict World meant that agents pursued shifting goals in reaction to their experiences as well as changing motivations in response to these experiences and available options (Heylighen 2016; Dorigo et al. 2000). Put simply, we found that the existence of a stable human nature and stable social system was itself an assumption that we should not have made. Making this assumption led to significant

misinterpretation of data and a poor understanding of the world in question, despite success in some tasks.

This finding lends novel support to the idea that even certain *forms* of social science inquiry (e.g., identifying utility maximizing behavior) can be effectively blind to coherent, rule-based behavior if agents' understanding of their world differs from researchers' understanding. Such an idea has gained prominence in discussions about systemic inequality, especially racial inequality, where scholars dispute what constitutes agentic versus structural sources of inequality (Royce 2018; Wilson 1987; Massey and Denton 1993). Our experience with the GT project indicates it is possible for agents to be embedded in a social structure perfectly navigable to themselves but so alien to scholars that the latter cannot conceive its existence. Moreover, the social ontology of Conflict World was relatively intuitive once revealed, suggesting that access to agents' own understanding of their social system (e.g., through interviews, ethnography, etc.) would have been key to understanding it correctly and thus making accurate inferences.

1.5 Data collection versus data generating processes: the language of observation

A final constraint that applied across simulated worlds was, again, the role of language, but this time in the context of data collection. Research teams requested information from simulation teams in order to test hypotheses about each simulation. These requests were a focal point of frustration due to (a) the desire to test the full range of social science research methods for studying simulated social systems and (b) the reliance of research teams on the resulting data to both construct a basic understanding of each simulation and complete GT tasks. The former is best illustrated by persistent attempts to employ qualitative methods, in particular ethnography. We suggest that such dissonance between stated GT project goals and the specific goals of our research team reflect distinctions between data collection as often practiced in the social sciences and data generating processes foundational to simulated social systems.

In short, it was effectively impossible to conduct ethnography in these simulated worlds. This finding was not self-evident at the time, nor do we believe it intentional on the part of simulation teams or the T&E team. Rather, the natural value of ethnography is its ability to leverage human perception to identify gaps in our understanding of social phenomena (Becker et al. 2004; Jessor et al. 1996; Small 2009; Pacewicz 2020), often through discovery of persistent, multi-dimensional configurations. For example, in Urban World if an agent enters a site and we can determine how that agent knows what time of day it is then we may better understand why agents spend less time at one type of site compared to another type of site—an important factor in human behavior exploited by casino designers. An ethnographic account of agents visiting each type of site could reveal that one type of site always has a clock prominently displayed on the wall while the other type of site

never does. However, we ultimately learned that agents simply “know” what time it is, and that time spent at sites was an inherent property of site type.

Similar instances of perfect information and just-so features of each world could be found in other simulations. For example, in Disaster World, hurricane dynamics were governed by fixed parameters with deterministic impact on the risk posed to regions/individuals. Yet our experience with this ontological question (i.e. “What observations are possible?” regardless of their truth value) provided an important lesson learned that complemented the lesson learned from Conflict World’s radically different ontology compared to the other simulations:

Regardless of the type of social system being studied, the accessibility of one type of data does not negate the need for access to additional types. Because we knew the simulations were simulations, we understood that they lacked the complexity of real life, but uncertainty around what was observable compounded uncertainty around the objects and causal relationships being simulated. This doubly shifting landscape is a common feature of real social systems and represents the social scientific equivalent of the “state of nature” where one has no knowledge of what is or is not socially meaningful. Historically, empirical social science was more qualitative in nature and made extensive use of interviews, participant observation, and ethnography. Over time, quantitative social science developed in parallel, emerging as soon as suitable methods were developed (Hacking 1990). Our experience with the GT project suggests that the initial use of qualitative methods as a tool for uncovering ontological properties of social systems logically precedes quantification of those properties.

The GT simulated worlds represented an anachronistic case where we began with quantification then attempted to understand how a social system works in order to develop effective interventions. However, we did not have the benefit of human perception to identify meaningful properties and/or their configurations in each world. Ultimately, this parallels the hyper-quantification of social behavior we observe today (Lazer et al. 2009; Edelman et al. 2020) and suggests that without the ability to observe the social system *in vivo* we risk developing a working model of behavior that excludes key properties of the world. For example, in Urban World there existed an entire process for “eating” that we never uncovered because we thought it was self-evident from other behavioral data, but this process proved crucial for understanding simulated disease transmission in Phase 3.

1.6 Ontological lessons learned

In addition to the specific lessons learned noted throughout, we identified two broad lessons learned from the ontological questions raised by the GT project. First, successful quantitative social science requires well-posed questions that use well-defined terms. Data does not, in itself, induce understanding, and descriptions of data can impart significant bias, even when it is known that such bias exists and could substantively affect analysis. Second, collective problem solving without a shared ontology about the object of study is extraordinarily difficult. Both have real

world implications, but we will focus briefly on the latter as the former may be an artifact of the GT project's design.

The lack of a shared ontology concerning what objects exist in a social system and the causal relationships between these objects generated significant challenges for resource allocation. Specifically, cognitive responses to GT tasks (i.e., problem solving) required some level of certainty concerning this ontology, but without this certainty many team members could not intuit how to proceed. For example, consistent lack of progress in learning how agents behaved in the Conflict World simulation made it difficult to justify devoting additional resources to associated tasks. Conversely, the expansive nature of the Urban World simulation meant that it was difficult to gauge whether progress had been made at all since we never knew if we had uncovered all relevant properties of the world. Both situations led to ambiguity about the entities and relationships under study which often resulted in pursuing an exhaustive understanding of causal relationships (i.e., perfecting the "Explain" task during each phase). However, this ambiguity has implications for real world problem solving in the form of issue advocacy and public policy recommendations.

A limited understanding of salient social entities (e.g., agents, processes, contexts) limited our ability to conduct experiments within these worlds. Experimentation requires control in the form of sufficient knowledge about relevant forces shaping behavior and the variation of one or a few factors. Without this knowledge, experiments can become fishing expeditions where other research methods may produce more insight with less effort. It was not until we approached the end of the project that we came to understand enough about these worlds to cultivate expectations about which features would be most insightful to systematically vary for the direct identification of causes and useful policies. This delayed understanding, combined with the significant turnaround time of research requests, made experimentation a high-risk endeavor during most of the project. Future simulation studies of this kind might consider how to support a sufficient ontological framework for posing meaningful experiments.

The GT project constituted a highly structured yet cooperative team-based assessment of agent based simulations as a platform for productively testing different methods for studying social behavior. Despite prolific uncertainty around the form and content of these simulated social systems, project participants committed to achieving the same goal under the belief that this goal was, generally speaking, beneficial to advancing social science and its applications. Most complex policy debates involve similar levels of uncertainty about the true nature of social problems and policymakers must typically make decisions based on an incomplete understanding of these problems (Head 2019; Ney 2009; Yung et al. 2019). However, policymakers are also often subject to a barrage of information from issue advocates, popularly termed lobbyists (Bok 2001; Nelson and Yackee 2012). These advocates advance a particular understanding of a social problem so that policymakers can address this problem in the manner they believe most effective.

Yet it is rarely clear which policies will be most effective because advocates must begin with their own assumptions about the social system they intend to influence (Markusen and Venables 1988; Schneider and Ingram 1990; Pielke et al. 2008). Competing assumptions will naturally produce disparate policy solutions. Thus,

uncertainty regarding assumptions in real world policy debates, to some extent, mirrors the ontological uncertainty experienced by GT participants. In the case of the GT project, participants' commitment to rigorous scientific inquiry and shared striving toward a common goal *were not* enough to overcome the effects of this uncertainty in many cases. More broadly, we tentatively conclude that ontological differences in how issue advocates understand the same social problems will constrain the ability of policymakers to identify compromise solutions to these problems. Further, our experience with Conflict World suggests it is possible for good faith actors to be *incapable* of identifying key features of social systems if they rely on a limited array of evidentiary sources.

It is unclear how this issue might be addressed in the short term except to acknowledge its effects. Our experience with the GT project, however, suggests that long term efforts to formalize ontological assumptions about social systems must be supported to avoid inefficient, ineffective, or counterproductive public policies enacted by elected politicians. Because there is much we do not understand about social systems, and representative political systems rely on popular understandings of social problems and their solutions, it is important that social scientists highlight this barrier to consensus building and begin adopting a means of communicating the social ontologies used within their work.

2 Epistemology

All work that overlaps neighboring fields, such as we occasionally undertake and which the sociologists must necessarily undertake again and again, is burdened with the resigned realization that at best one provides the specialist with useful questions upon which [they] would not so easily hit from [their] own specialized point of view.

~ *Science as a Vocation*, Weber (1958)

The type and extent of epistemological issues associated with the GT project were often predictable and provided significant insight regarding fundamental tensions associated with (a) doing social science research versus (b) applying findings from social science research. Issues, such as choice of problem solving strategy for each task, stemmed directly from the structure and stated goal of the project. We do not discuss the specifics of these issues but, as with our examination of ontological issues, focus on broader questions with substantial implications for collective problem solving. These questions arose from the nature of the GT tasks performed during each phase of the project and speak to what Weber termed "science as a vocation" in reference to the external economic forces associated with and the lack of intrinsic value characteristic of scientific inquiry in practice. We thus organize our discussion of these issues and related lessons learned according to each type of GT task.

A clear pattern emerged over the course of the GT project around the types of knowledge that were necessary to complete GT tasks. Each phase of the project was identical insofar as research teams had to complete the same three tasks for each

world: explain, predict, and prescribe. While social systems differed in their structure and content according to the simulation teams' models, explaining each social system, predicting out-of-sample properties of each social system, and prescribing interventions for each social system represented loosely related yet distinct epistemological goals.

The first epistemological goal during each phase of the GT project was to explain the simulated social system. Explanation was defined as the development of a directed acyclic graph constituted by all relevant features (i.e., nodes) and causal relationships between features (i.e., directed edges). Relevancy was defined through prompts—such as “explain how agents form friendships”—that served to anchor inquiries for additional information. This GT task required coping with the ontological uncertainties noted above to produce an acyclic graph effectively reproducing the agent-based rules and/or agent-based model parameter relationships used to generate the simulation data provided to research teams.

A defining characteristic of this task was its holistic framing. While prompts provided a way to help research teams begin studying each phase's simulated world, research teams were evaluated according to their ability to uncover nodes and directed edges. This evaluation approach presents an egalitarian epistemology where all knowledge had approximately equal value regardless of the relative importance of individual pieces of knowledge for understanding the social system. This egalitarianism incentivized investigation of potentially second and third order effects that were, at best, tangentially related to properties of interest. Any knowledge was good knowledge.

A secondary characteristic of this task was significant slippage between what research teams believed they were investigating and what simulation teams had identified as relevant features. Specifically, the ground truth under study was an assortment of data generated by simulation teams and analyzed by research teams. The terminology used by simulation/research teams was, as noted, ill-defined and referred to simulated social processes. To “explain” these social systems research teams had to name causal nodes and provide a written explanation of how each node influences another if a directed edge existed. However, evaluating these explanations required simulation teams to interpret nodes and directed edges according to their internal understanding of the data generating processes they had developed.

This situation meant it was possible that research teams could describe a simulated social process but, without further clarification, simulation teams infer research teams were referring to a different simulated social process than intended, a phenomena reminiscent of boundary objects (Star and Griesemer 1989) that attract shared attention despite being understood and used in very different ways. Once the ground truth of each simulation was revealed, it was clear that there were many cases where this applied, such as agents' affinities for particular sites in Urban World. While our understanding of site visitation was based on the desire to form and maintain friendships with similar agents, the Phase 2 Urban World simulation replicated this agent-based form of homophily through a process where agents chose to visit preferred sites based on both agents' and sites' characteristics. However, our understanding of the ground truth causal diagram for Urban World's Phase 3 simulation suggests that

agents of a similar type formed friendships by choosing to visit similar sites rather than explicitly choosing similar agents. This minor distinction had serious implications for our ability to understand the simulated social system in the predict and prescribe tasks and highlights the fact that simulation/research teams could believe they are referring to the same social processes verbally (i.e., homophily in friendship formation) but are actually referring to different ground truths. Our experience with this epistemological issue suggests that clearly defined, shared referents are vital for constructing an accurate ground truth understanding of a social system, and we recommend future simulations testing the utility of different social science methodologies remove this ambiguity to ensure the robustness of results.

Conversely, terminological ambiguity also meant that research teams could refer to features using different words or phrases than simulation teams yet still reference the same ground truth processes. For example, it is unclear how much our failure to identify the nodes and directed edges associated with the process of food consumption in Urban World affected the evaluation of our explanation of this world in Phase 3. Without additional information, this lack of clarity cannot be resolved since we assumed data referring to relevant agent behavior (e.g., entering a restaurant) were sufficient to understand the process of quelling hunger and we could not know this explanation was deficient since we, as a research team, did not have access to the ground truth (i.e., the simulation) or other feedback indicating that our understanding was incomplete (i.e., knowledge about unobserved features). Our experience suggests it is vital to identify such measurement issues before attempting to construct an exhaustive causal model of a social system. This finding supports multiple arguments made in the social sciences about the role of measurement theory and highlights their applicability to data-driven analysis where measurement is assumed to be error free (or correctable with sufficient data) (Shultz et al. 2013; Bandalos 2018; Leplège 2003; Goertz and Mahoney 2012).

The second epistemological goal of each phase of the GT project was to predict out-of-sample characteristics of the simulated social system. For each world, simulation teams provided data for a discrete period of time. The primary goal of research teams was to then make predictions about what happens immediately following this time period. Evaluation metrics were typically based on predictions made over a discrete period of time (e.g., 7 days of simulated behavior). This GT task avoided many of the ontological issues noted above since the simulated data was itself the ground truth for each world and research teams did not need to explain the set of causal relationships used to generate predictions.

A defining characteristic of this task was a data-centric approach to modeling each simulated social system. Because the goal was to predict outcomes based on the data provided it was not necessary to construct any broader understanding of the social system than was needed to accurately predict future behavior. This feature of the predict task provided both natural scoping conditions for analysis and clear priorities in terms of analytic effort. The most efficient strategy was to identify the primary causal relationships governing the outcome of interest, such as the average number of friendships in the world or agents' responses to a natural disaster. Other aspects of the simulation could be ignored to the extent that they did not affect this outcome.

In some ways, the prediction task was the most familiar to our research team members as it most accurately reflects both current expectations around effective data science and historic efforts in the social sciences to identify the features most important for understanding social phenomena. For example, residential segregation is a well-known phenomenon in the United States which could allow simple and often correct guesses about your neighbors' race given your own (Massey and Denton 1988; Charles 2003; Reardon et al. 2015). Even if predicting your neighbors' race based on your own race is not something social scientists typically do, the fundamental epistemological issue remains reproducibility via prediction. In some ways, data scientists are more familiar with this issue given that many applications of machine learning involve generating out-of-sample predictions. Our experience with the GT project suggests that, when interpretation of features is irrelevant, even basic data science approaches, such as clustering, can yield a huge payoff in terms of predictive accuracy.

Notably, however, the explain task and predict task were only loosely related. Knowledge production in the former required a robust understanding of the social systems' ontology in order to exhaustively test possible causal relationships, but knowledge production in the latter only required a "good enough" understanding of each social system to accurately reproduce and then predict agent behavior. Emphasis on explicit directed causal relationships and predictive accuracy by those studying causal inference in the social sciences would suggest both tasks are important, but our experience with the GT project suggests the opposite: Fully understanding how a social system operates does not necessarily reveal which parts of this system are most important for outcomes of interest, and the ability to predict outcomes of interest does not indicate a full understanding of the social system. One can even imagine a GT task between explanation and prediction in which a causal weight is assigned to each edge in the graph.

There are three potential implications from this observation. First, if the goal of generating new knowledge is not clear, then it is easy to fall into a trap where researchers examine either a small part of a social system or attempt to fully explain the social system when one *or* the other will suffice. Second, marginal advances in our understanding of a social system need to be put into context relative to an outcome of interest. If no outcome is specified, then the value of these advances cannot be judged. Third, when the social system is unintuitive, as in the case of Conflict World, then it may be unclear whether a full or partial understanding has been attained. This last implication is the most serious as it means we can have a working understanding of one part (or even multiple parts) of a social system that produces good predictive accuracy but that does not capture key causal relationships operating essentially out of sight. Consider Captain Cook's fateful encounter with indigenous Hawaiians: he believed he understood enough about Hawaiian culture to play god, but was killed when he and his men failed to meet Hawaiian expectations about how gods behave. It was not that Cook could not predict the average reactions of the Hawaiians, but that he did not fully understand the social system generating those reactions (Sahlins 1995).

The third and final epistemological goal of each phase of the GT project was to optimize a set of policy prescriptions to maximize or minimize an outcome of

interest. As with the predict task, research teams were evaluated based on out-of-sample prescriptions where interventions had been introduced into the simulation as prescribed. The primary goal of research teams was to construct counterfactual predictions of agent behavior based on their understanding of how each world operated. In this respect, the predict and prescribe tasks were tightly coupled, largely avoided ontological issues noted above, and reflected a pragmatic approach to knowledge production. The prescribe task differed from the predict task in that the effects of interventions had to be estimated and, once known, optimized to elicit the “best” possible outcome in each world. This task was akin to data-driven policymaking where the effectiveness of policies are empirically tested after the fact. Given the practical implications for policymakers, the purpose of this knowledge production was clear and the problem (i.e., optimization) well-defined. As with the predict task, however, only a pragmatic understanding of each social system was required to generate high quality prescriptions.

Our experience with the prescribe task during each phase of the GT project again led us to conclude that the types of knowledge generated during each phase were loosely coupled, at best. Identifying the optimal timing and implementation of prescriptions for each social system did not require a complete understanding of its causal structure. It also did not necessarily require high predictive accuracy, only high impact prescriptions leading to the best outcomes possible. The latter was illustrated in Urban World during Phase 3 when efforts to develop an accurate predictive model of the simulation fell short of our expectations, yet we had sufficient knowledge of the world to develop effective prescriptions based on little more than logic and basic estimates regarding the relative efficacy of each type of prescription available.

Other epistemological issues came to the fore during the GT project, such as what constituted ground truth and how to evaluate it, but we have organized our experience by task since each task had unique requirements that led to fairly discrete forms of knowledge production. Performing well on each task required a slightly different type of knowledge. Due to the fast-paced nature of the GT project, these differences manifested in our problem solving (i.e., knowledge production) strategies. As noted, the epistemological approaches used for each task were only loosely related: explaining the social system had marginal benefit for predicting behavior, and predicting behavior had marginal benefit for developing an optimal set of prescriptions. While fine-grained causal information mostly distracted from the largest factors impacting outcomes of interest, knowing the largest factors impacting outcomes of interest mostly distracted from the relative impact of each potential intervention and thus the task of formulating an optimal portfolio of interventions.

Our experience with epistemological issues during the GT project suggests three possible lessons learned of use to an array of stakeholders ranging from non-profit organizations to academic researchers to public office holders. First, understanding a social system is not the same as learning about a social system. It is possible to learn more than enough about a social system to accurately predict behavior and generate effective policy prescriptions without a thorough understanding of the system as a whole. Yet this pragmatic focus, which creates a natural scope condition for data collection and analysis, can miss pivotal features of the system while still

performing well on predictive/prescriptive tasks. Our tentative conclusion is that the need to balance holistic understanding with pragmatic inquiry emerges precisely because the knowledge produced by each strategy speaks to different epistemological goals. Alternatively, it is possible that real world social systems exhibit more tightly coupled relationships between these goals, as in cases of heterogeneous treatment effects that imply a more complex set of causal relationships than originally assumed when designing an intervention.

A second lesson learned was that knowledge production in an academic setting, specifically quantitative research in the social sciences, often, but not always, focuses on marginal improvements in our understanding. This focus means an academic mode of inquiry, which involves the demonstration of new social objects, forces, or relationships, will tend to be less effective at generating policy relevant findings. It is not that scholars lack the tools to produce such findings, but that the epistemological goal of social science differs from that of policymaking. However, pragmatic and holistic approaches to policy development and implementation are not antithetical to the practice of academic research insofar as scholars can demonstrate that their marginal improvement in our understanding has a substantive effect on outcomes of interest. In fact, this approach is often adopted as a model for evidence-based policymaking and major grantmaking organizations often demand that proposals explicitly identify concrete implications for policy (e.g., through requiring randomized clinical trials of interventions).

A final lesson learned was that, even if knowledge produced by quantitative academic researchers in the social sciences has identified the primary causal relationships associated with an outcome of interest, it is very difficult to uncover this information for reuse. To find and apply this work requires not just methodological competency (e.g., ability to distinguish between high quality and low quality studies) but also (a) domain knowledge of the field/sub-field and its internal debates, (b) willingness to identify and attempt to overcome disciplinary biases both in the field/sub-field and as individuals, and (c) access to diverse sources of research, including not just a wide range of peer-reviewed journals but also respected, if informal, repositories for research, such as the National Bureau of Economic Research (NBER) working papers, and high quality studies produced by non-academic institutions run by academic researchers. The amount of time and energy necessary to effectively search the literature is thus prohibitive even for the best trained, most well-read social scientist, suggesting that devolving that burden onto policymakers, which is often the tacit strategy of social science, is unrealistic. In fact, our experience with the GT project suggests that the inability for social scientists to move beyond their own disciplinary biases can lead to significant wasted time and effort when those biases strongly suggest a course of action inapplicable to the situation at hand.

3 Methodology

With the ontological and epistemological issues above in mind, we now turn our focus to methodology. It was expected that methodological issues would be both prolific and highly productive. We found this to be true but not always in ways we anticipated. The variety of approaches employed reflected a diversity of perspectives about which methods might be most applicable in each world during each phase and task. While not always followed, we found the strategy of working backward from the goal of each task helpful for constraining the set of methods to be tested. When we found a particular method was useful in one world/phase/task we often attempted to deduce its relevance in another world/phase/task. Before discussing an example of each strategy, it is useful to review both the variety of methods applied and the immediate methodological issues they often raised.

During the GT project, members of our team applied a menagerie of methods ranging from information visualization and bespoke Google queries (e.g., number of webpages on which a pair of Conflict World labels appeared) to standard statistical models to machine learning approaches for agent-based models of complex systems to emerging artificial intelligence techniques to simple searches of the Internet. These included geographic information visualization, correlation, linear regression, logistic regression, Shapley regression, Granger causality estimation, auto-regression and time series analysis, sparse regression, survival models, markov models, clustering, Bayesian graphical models, decision trees and random forests, ensemble methods (e.g., boosting, bagging), support vector machines, k-nearest neighbors, Hawkes process analysis and simulation of spatially interdependent point processes with probabilistic finite-state machines, agent-based models, probabilistic programming models (Wood et al. 2014), along with a wide array of deep learning approaches from standard, feed-forward artificial neural networks (Fine 2006) and recurrent neural networks (Rumelhart et al. 1986), to auto-encoders (Hinton and Salakhutdinov 2006), sequence-to-sequence neural networks (Sriram et al. 2017), graph convolutional networks (Kipf and Welling 2016), theory of mind neural networks (Rabinowitz et al. 2018), and world models (Ha and Schmidhuber 2018). We also used network analysis (Wasserman and Faust 1994) and methods for analysis of GIS data when applicable. Choice of method was, ideally, based on its suitability to a particular task and/or form of data. We first review general approaches to method selection and their efficacy before discussing three approaches that showed the most promise—Granger causal networks, probabilistic programming, and neural networks.

Given the varied backgrounds of team members, bias towards particular habits in problem solving sometimes led to (over)-reliance on a favored method. For example, regression analysis and clustering algorithms were commonly employed if only because they were familiar tools for exploring unfamiliar data. This meant that choice of method was not always ideal given the task/data, but the relative suitability of a method would often quickly become apparent. One clear case of this situation was when we attempted to study a social network in Urban World with more than 1000 nodes using exponential random graph models (ERGMs) only to find one

of the most robust packages for studying ERGMs struggled to estimate even simple properties of the network (Handcock et al. 2008). Conversely, there were times when clustering algorithms outperformed more refined models based on our working understanding of the social system. In this regard, pragmatism was the rule and deliberate planning the exception.

Related to bias in habits of problem solving, domain specific knowledge affected both the collection and interpretation of data. Each world tended to encompass an overarching set of social processes that could be described in succinct and coherent terms. For example, we came to understand the first simulation as Urban World because it involved agents living in a geographic space and moving between points within this space to achieve a variety of goals, in the process entering and exiting sites where they performed actions according to site type. This conception arose out of our team's intuition about the features within the simulation and their relationship to each other. In response, analysis of Urban World was primarily left to an expert in GIS methods and an urban sociologist. However, the Urban World simulation team was primarily composed of geographers.

The result was that Urban World was aptly named, and this heuristic often helped us to intuit the existence of basic features, but that differences in domain expertise between geography and sociology led to focus on different social processes. This disparity was most evident in the fact that agents did not directly pursue relationships with similar agents but formed relationships with similar agents by seeking out sites of shared interest where they could then meet and, potentially, form a friendship. The former is foundational in sociological understandings of friendship formation (McPherson et al. 2001), while the latter is foundational to geographic understandings of travel patterns (Liu et al. 2015). To grossly exaggerate this distinction, we might say those who built "Urban World" viewed it as a social system where people often focused on the process of choosing where to go next while our team viewed it as a social system where people often focused on the process of choosing friends. Misalignment in domain expertise, in this case, taught us an important lesson: Similar understandings of a social system can be equally grounded in empirical research but methodological strategies aligned with the ground truth (i.e. a geographic perspective on urban social systems) may better catalyze knowledge production.

This illustrates and underscores the No Free Lunch theorems in machine learning proved by David Wolpert and William Macready (Wolpert and Macready 1997), which "state that any two optimization algorithms are equivalent when their performance is averaged across all possible problems" (Wolpert and Macready 2005)—or all possible worlds. If a method works well in some social world, it will work poorly in another. Our understanding and methodological strategies worked well when they aligned, and poorly when they did not.

An alternative to applying domain specific knowledge to study a social system is to adopt a Bayesian approach to knowledge production. Specifically, domain expertise generates strong priors about the form and function of a social system, but it is always possible to consider semi-informative priors to tentatively adopt then test. A significant amount of basic analysis involved this kind of work. This often meant

iterative testing of relationships between features of each social system with basic constraints on what relationships were thought to be possible. We found that this could produce results where all features were related to all other features but, when the approach was properly specified, could also help advance our understanding of each world significantly.

While making too many assumptions could prove problematic and systematically testing a set of assumptions could prove useful, a third methodological strategy was to begin from the *tabula rasa* of no (or as few as possible) assumptions about each social system. Methods that embody this strategy were highly data driven, which translated into a need for high levels of technical expertise unrelated to social science. Under the circumstances, these methods often showed the most promise for use in social science research precisely because they were not beholden to relationships we expected to find and focused on uncovering or modeling the relationships we *did* find in the data. This strategy was especially useful given the ontological issues we faced.

For example, we explored whether we could construct causal predictors from observation sequences of variables alone. Designing an efficient causality test, that may be carried out in the absence of restrictive presuppositions on the underlying dynamical structure of the data at hand, is non-trivial. Nevertheless, ability to computationally infer statistical *prima facie* evidence of causal dependence may yield a far more discriminative tool for data analysis compared to the calculation of simple correlations. On this line of thought, we devised a non-parametric test of Granger causality for quantized data streams realized from the variations of the observed variables in the world simulations.

In contrast to state-of-the-art binary tests, this approach computes the degree of causal dependence between data streams, without making any restrictive assumptions, linearity or otherwise. Additionally, without any a priori imposition of specific dynamical structure, we were able to infer explicit generative models of causal cross-dependence, which may then be used for prediction. These explicit models are represented as generalized probabilistic automata, referred to crossed automata, and are shown to be sufficient to capture a fairly general class of causal dependence (Chattopadhyay 2014). The proposed algorithms are computationally efficient in the probably approximately correct (PAC) sense (Valiant 1984); i.e., we find good models of cross-dependence with high probability, with polynomial run-times and sample complexities. The causality network inferred from this dataset revealed non-trivial relationships, and laid the groundwork for such deep data-driven interrogation of complex social phenomena in the future, particularly in situations where sequential observations on many interacting variables are available.

Another data driven approach that enabled the explicit incorporation of social theoretical intuition was probabilistic programming (e.g. Salvatier et al. 2016). Probabilistic programming languages (PPLs) allow stochastic elements to be included in deterministic models by treating statistical distributions as objects on which we may perform basic logical/mathematical operations. This functionality further allows us to create a generative model of behavior within which we may embed prior information about the social system (Goodman et al. 2012). For example, a Bernoulli random variable, such as a coin toss turning up heads, may determine whether a person

in Urban World contracts a disease given exposure, and the severity of the disease for that person represents a second random variable, perhaps normally distributed such that it typically causes discomfort and the potential to infect others but not lasting disability. In extreme cases, however, the disease may cause death, or spread much more rapidly than normal, like COVID-19 at a super spreader event (Althouse et al. 2020). Within a PPL framework, the coin toss determines whether the disease is passed then contraction itself has a chance (e.g., based on a different statistical distribution) of leading to transmission and loss of health or loss of life. If we wanted to understand the effect of contraction on severity and transmission, we would observe the distribution of these events, conditional on one another, then tune the probabilities of our program in order to generate the appropriate distribution of outcomes, which would later be available for us to determine whether or not a new disease had emerged from the same “world” as the last.

Within the GT project, early stages of Power World used PPL as a means to answer questions about patterns in the data, trying to support or disprove potential hypotheses about the way the world works (e.g., “Does the team with the highest productivity always win conflicts?”). A sketch of the overarching structure of the world was built using information provided in briefings and communiques. We knew the general outline of the program we were modeling but only general things about what happened at distinct points in time. We hard-coded the things we did know and set a parameterized distribution over the space of programs consistent with behavior expected at the unknown regions. PPLs are able to search this space of programs (i.e., potential worlds) to find the one most consistent with observed data.

Many of the hidden processes we wanted to model, such as conflict between two groups, had binary outcomes. To figure out which features contributed to a given outcome, we mapped each state of each feature to a value and then used linear combinations of these values to produce the weighting of a coin, the flip of which stood-in for the process we were modelling. This approach of searching through the space of weights for factors helped guide data analysis. For example, if the inferred program for determining outcomes for group conflicts weighed group sizes heavily, then we would know to try to look at the data to see if indeed a large group size was reported shortly before a conflict and whether it was correlated with victory. Once features had been narrowed down to those contributing most strongly to certain outcomes, we could hand-craft competing models that only considered those features.

PPLs iterate over a “program trace” (Cusumano-Towner et al. 2019) or snapshot of various states during the execution of a probabilistic program. However, our program was written such that it could also provide us with a likelihood of actually observing a particular trace. Given that we know that at a particular time the target system was in a particular state, we could force our program to make those same choices and change any other unconstrained choices in our program such that it maximized the likelihood of the observed data. One such choice might be the weight given to a particular feature.

The challenge arises when trying to propose a new value for a particular choice, especially when choices are tightly coupled. If a change we proposed to a variable made the trace less likely we were cautious about accepting it. One simple but common problem occurred when variable A only takes on a particular state when the

states of variables B and C agree (i.e., correlate). It is possible to detect such cases and handle them appropriately with PPLs, but it currently requires explicit knowledge of the underlying system as well as expert-level understanding of the MCMC algorithms employed. Furthermore, under certain conditions, probabilistic programs are guaranteed to converge, but they are not guaranteed to converge quickly. Reasonable convergence time comes down to well-designed model spaces informed by knowledge of the target system. In this regard, PPLs may be well-suited to doing social science in domains where experts are readily available, well-informed, and forthcoming about plausible and implausible mechanisms of behavior.

A final promising methodological approach we found for studying agent behavior was deep learning in general, and graph convolutional networks (GCNs) in particular (Kipf and Welling 2016). GCNs are a subset of graph neural network (GNN) models that characterize nodes by including features of neighboring nodes (Wu et al. 2020). We also tested the efficacy of attention-based approaches, weighting neighboring nodes according to their “importance” for ego nodes. To capture the time-varying nature of the networks involved we explored the use of long short-term memory (LSTM) propagation for constructing successive GCNs over time (Hochreiter and Schmidhuber 1997), one of many ways GNNs can be constructed to suit specific use cases (Zhou et al. 2018). Finally, we tested an approach to dynamic graphs (EvolveGCN) designed to reduce computational complexity by focusing on temporal dynamics over node representations (Pareja et al. 2020) and the use of hyperbolic GCNs that preserve scale-free or hierarchical graph structures (Chami et al. 2019). The most salient use case for these approaches was Urban World, where all three phases/simulations included multiple dynamic social networks in the form of friendships, work relationships, and site co-location.

In essence, GNN methods assign latent states to graph nodes by embedding these nodes in a geometric space. GCNs assign latent states based on the states of a node’s neighbours (identified from an adjacency matrix) and can be propagated using an LSTM-like mechanism. Indeed the aforementioned methods seem to adequately capture the structure of social science problems in question: Agents can be represented by nodes, their associations by edges, and time evolution corresponds to the evolution of agents’ states. These approaches seemed a natural fit during, for example, Phase 3 of Urban World where tasks focused on understanding, predicting, and intervening in disease transmission networks. For instance, we expected that the embedding of agent features using latent states would prevent human bias in feature selection while still retaining maximal information.

However, these methods only partially solved the problem of disease evolution because almost all focus either on node state prediction or link prediction separately. Only the EvolveGCN framework purported the ability to do both simultaneously, but this feature was novel and its implementation brittle. In the case of these synthetic worlds, many types/levels of associations existed: node states had discrete properties, and we had to predict the evolution of the whole system, not just the state of nodes, for example. All of these are complications we believe had yet to be addressed by methods at the time but represent ready targets for the future.

Perhaps the most notable aspect of applying GCNs in the GT project was their persistent inability to reproduce macrosocial properties of social systems based on

microdata. That is, disease transmission is a fundamentally network-based process and so we expected GCNs to perform well when modeling the variety of factors influencing this process. However, we were never able to reproduce the accuracy obtained from applying simple compartmental models of disease evolution, such as the susceptible, infected, and recovered (SIR) model. These models consist of basic differential equations whose parameters determine the overall distribution of disease states within the population at an aggregate level. We expected that GCNs would at least be able to reproduce (if not improve on) compartmental models but, despite significant effort, we found that they could not. Given the structural similarity between the GCN architecture and network-based mechanisms of disease transmission we are forced to conclude that either (a) our implementation of GCN was conceptually flawed, or (b) the use of GCN fails to capture a fundamental property of network evolution. We discovered later that disease spread in Urban World was modeled with SIR-like models and so our models may have been too precise for the coarse-grained spread of disease in data. In either case, GNNs (and especially GCNs) appear to be a promising new method for network analysis in the social sciences but may require further development before scholars can realize their full potential in complex social settings.

The methodological issues we faced during the GT project suggested three basic lessons learned. First, the level of analysis and type of social process involved are critical for selecting the appropriate method. This lesson is almost remedial in nature given that it amounts to a reminder to select the right tool for the right job. However, the second lesson was methods that may seem intuitively applicable can fail spectacularly (e.g. GCNs), but that openness to alternative approaches can allow for a process of self-correction. Sometimes the latest and greatest method seems like it *should* work but does not, and that failure to perform as expected can be useful for thinking about less complex but similarly applicable methods with a proven track record and which still embed unarticulated understandings about the world in question. Finally, we learned that imperfect knowledge about a social system can be good enough to find effective methods. Beginning from a *tabula rasa* typically does not imply beginning from a state of total ignorance. Rather, acknowledging some level of ignorance can help guide the use of methods that have few if any assumptions and may thus “enlighten” our thinking about a problem.

4 Discussion

The ambitious DARPA Ground Truth project led to the simulation of four social worlds in which social science could be evaluated *in silico*. Because these worlds were based on simulations, simulation teams knew the causal ground truth—they had designed the programs themselves—but the research teams did not. Our experience attempting to crack puzzles of these worlds reinforced what AI pioneer Allen Newell stated about research: “You can’t play 20 questions with nature and win” (Newell et al. 1972). And we couldn’t play 20 questions about *in silico* social worlds and win consistently. Stochastic elements of the simulations resulted in a Bayes error rate or irreducible error far greater than 0, and natural limitations on certain

forms of data gathering like ethnography and other qualitative methods in the *in silico* setting were awkward and limited the context research teams were able to achieve. Nevertheless, we and the other research teams were able to do better than random change on most tasks in most phases, and we improved over time and with additional data. Moreover, by confronting tasks with distinct ontological, epistemic and methodological requirements, we gained deep insight into the limits of quantitative social science, especially with respect to informing social policy.

Faced with unfamiliar simulated worlds, we struggled to identify their underlying ontology. This highlighted the crucial role of grounded, qualitative insight from insider views of any social system, which cannot not be substituted with quantitative censuses or digital trace data. Why? Because data labels did not provide enough context. They became boundary objects, passed from simulators to researchers through T&E without a shared certainty of reference. This was not a flawed property of the GT program but reflects the limits of ungrounded quantitative social science—data science—where variable names disseminate with interpretations that shift with context.

Without a tighter sense of not only the ontology of GT worlds, but what was salient, we struggled to construct experiments despite their availability as a sanctioned data gathering approach because, until the end, we did not know which critical factors to vary, holding others constant. This underscored the challenges of problem solving under conditions of extreme existential uncertainty that contribute to many complex societal challenges. The policy relevance of quantitative social science is also conspired against by the current epistemic standard for publication. Demonstration of novel entities and causes is expected in science, but this narrow exhibition can work against the ability to make meaningful interventions on problems and propose robust policies—from above or below.

Finally, we attempted to use a vast menagerie of methods. Some of the most promising emerging methods included detailed bespoke descriptive data analysis, probabilistic programming, deep neural networks of many flavors, and systems of predictive probabilistic finite state machines, which we developed alongside robust statistical and machine learning approaches to supervised and unsupervised learning. Through this exploration, we learned that imperfect knowledge about the most important factors can be sufficient to generate robust predictions and policies. Moreover, applying competing approaches via distinct subteams, including at one point the vast TopCoder.com global community of program solvers, enabled us to discover relevant structure underlying worlds that singular investigators and methods could not.

Collectively, these lessons suggest how different a policy-oriented quantitative social science would be than the quantitative social science and data science most commonly practiced to date. Data science and quantitative social science that serves policy will need to endure more failure, sustain more diversity, tolerate more uncertainty, and allow for more complexity than current institutions are well-positioned to support.

Acknowledgements The authors gratefully acknowledge DARPA grant HR00111820006 for the Ground Truth program, for Adam Russell, the architect of that program, and other participants in the program

(and authors of articles in this special issue) for their inspiration as fellow travelers and contributors to this project.

References

- Althouse BM, Wenger EA, Miller JC, Scarpino SV, Allard A, Hébert-Dufresne L, and Hao H (2020) "Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control." *PLoS Biol* 18(11): e3000897. <https://doi.org/10.1371/journal.pbio.3000897>
- Bandalos DL (2018) *Measurement theory and applications for the social sciences*. Guilford Publications, New York
- Becker HS, Gans HJ, Newman KS, Vaughan D (2004) On the value of ethnography: sociology and public policy: a dialogue. *Ann Am Acad Pol Soc Sci* 595(1):264–276
- Bok DC (2001) *The trouble with government*. Harvard University Press, Cambridge
- Chami I, Ying R, Ré C, Leskovec J (2019) Hyperbolic graph convolutional neural networks. *Adv Neural Inf Process Syst* 32(December):4869–4880
- Charles CZ (2003) The dynamics of racial residential segregation. *Ann Rev Soc* 29(1):167–207. <https://doi.org/10.1146/annurev.soc.29.010202.100002>
- Chattopadhyay, I. 2014. "Causality networks." *arXiv [cs.LG]*. arXiv. <http://arxiv.org/abs/1406.6651>. Accessed 24 Nov 2022
- Cheng J, Greiner R, Kelly J, Bell D, Liu W (2002) Learning Bayesian networks from data: an information-theory based approach. *Artif Intell* 137(1):43–90
- Cusumano-Towner, MF, Feras AS, Alexander KL, and Vikash KM. (2019). "Gen: a general-purpose probabilistic programming system with programmable inference." In: *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 221–36. PLDI 2019. New York, NY, USA: Association for Computing Machinery.
- Dahrendorf R (1973) *Homo sociologus*. Routledge Kegan Paul, London
- Denzinger J. 1995. "Knowledge-based distributed search using teamwork." In V Lesser (ed), *Proceedings of the First International Conference on Multiagent Systems*, 81–88. Association for the Advancement of Artificial Intelligence.
- Dorigo M, Bonabeau E, Theraulaz G (2000) Ant algorithms and stigmergy. *Futu Gener Comput Syst: FGCS* 16(8):851–871
- Edelmann A, Wolff T, Montagne D, Bail CA (2020) Computational social science and sociology. *Ann Rev Sociol* 46(1):61–81
- Fine TL (2006) *Feedforward neural network methodology*. Springer Science & Business Media, Heidelberg
- Goertz G, Mahoney J (2012) Concepts and measurement: ontology and epistemology. *Soc Sci Inf. Information Sur Les Sciences Sociales* 51(2):205–16
- Goodman N, Vikash M, Daniel MR, Keith B, and Joshua BT. (2012). "Church: a language for generative models." *arXiv [cs.PL]*. arXiv. <http://arxiv.org/abs/1206.3255>. Accessed 24 Nov 2022.
- Granovetter M (1985) Economic action and social structure: the problem of embeddedness. *Am J Soc* 91(3):481–510
- Hacking I (1990) *The taming of chance*. Cambridge University Press, Cambridge
- Ha D, Schmidhuber J (2018) Recurrent world models facilitate policy evolution. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) *Advances in neural information processing systems* 31. Curran Associates Inc, New York, pp 2450–62
- Hall KL, Vogel AL, Huang GC, Serrano KJ, Rice EL, Tsakraklides SP, Fiore SM (2018) The science of team science: a review of the empirical evidence and research gaps on collaboration in science. *Am Psychol* 73(4):532–548
- Handcock MS, Hunter DR, Butts CT, Goodreau SM, Morris M (2008) Statnet: software tools for the representation, visualization, analysis and simulation of network data. *J Stat Softw* 24(1):1548
- Head BW (2019) Forty years of wicked problems literature: forging closer links to policy studies. *Policy Soc* 38(2):180–197. <https://doi.org/10.1080/14494035.2018.1488797>
- Heylighen F (2016) Stigmergy as a universal coordination mechanism I: definition and components. *Cogn Syst Res* 38(June):4–13

- Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507
- Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
- Holzhauser S, Krebs F, Ernst A (2013) Considering baseline homophily when generating spatial social networks for agent-based modelling. *Comput Math Organ Theory* 19(2):128–150
- Jessor R, Colby A, Shweder RA (1996) *Ethnography and human development: context and meaning in social inquiry*. University of Chicago Press, Chicago
- Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. *ICLR* 2017. <https://openreview.net/forum?id=SJU4ayYgl>. Accessed 24 Nov 2022.
- Lazer D, Pentland A, Adamic L, Aral S, Barabasi A-L, Brewer D, Christakis N et al (2009) Social science. Computational social science. *Science* 323(5915):721–723
- Leplège Alain (2003) Editorial. Epistemology of measurement in the social sciences: historical and contemporary perspectives. *Soc Sci Inf. Information Sur Les Sciences Sociales* 42(4):451–62
- Li T, Yi H, James E, and Ishanu C. 2019. “Long-range event-level prediction and response simulation for urban crime and global terrorism with granger networks.” *arXiv [stat.AP]*. arXiv. <http://arxiv.org/abs/1911.05647>. Accessed 24 Nov 2022.
- Liu Xi, Gong Li, Gong Y, Liu Yu (2015) Revealing travel patterns and city structure with taxi trip data. *J Transp Geogr* 43(February):78–90
- Markusen JR, Venables AJ (1988) Trade policy with increasing returns and imperfect competition: contradictory results from competing assumptions. *J Int Econ* 24(3):299–316. [https://doi.org/10.1016/0022-1996\(88\)90039-6](https://doi.org/10.1016/0022-1996(88)90039-6)
- DS Massey and NA Denton (1993) *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA
- Massey DS, Denton NA (1988) The dimensions of residential segregation. *Soc Forces* 67(2):281. <https://doi.org/10.2307/2579183>
- McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. *Ann Rev Sociol* 27(1):415–444
- Nelson D, Yackee SW (2012) Lobbying coalitions and government policy change: an analysis of federal agency rulemaking. *J Politics* 74(2):339–353. <https://doi.org/10.1017/S0022381611001599>
- Newell A, Simon HA et al (1972) *Human problem solving*, vol 104. Prentice-Hall Englewood Cliffs, NJ
- Ney S (2009) *Resolving messy policy problems: handling conflict in environmental, transport, health and ageing policy*. Routledge, London. <https://doi.org/10.4324/9781849772389>.
- Pacewicz Josh (2020) What can you do with a single case? How to think about ethnographic case selection like a historical sociologist. *Sociol Methods Res.* <https://doi.org/10.1177/0049124119901213>
- Padgett JF, Powell WW (2012) *The emergence of organizations and markets*. Princeton University Press, New Jersey
- Pareja, A, Domeniconi, G, Chen, J, Ma, T, Suzumura, T, Kanezashi, H, Kaler, T, Schardl, T, and C Leiserson (2020) EvolveGCN: Evolving Graph Convolutional Networks for Dynamic Graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04), 5363–5370. <https://doi.org/10.1609/aaai.v34i04.5984>
- Pielke R, Wigley T, Green C (2008) Dangerous assumptions. *Nature* 452(7187):531–532. <https://doi.org/10.1038/452531a>
- Rabinowitz NC, Perbet F, Song HF, Zhang C, Ali Eslami SM, Botvinick M (2018) Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80:4218–4227
- Reardon SF, Fox L, Townsend J (2015) Neighborhood income composition by household race and income, 1990–2009. *Ann Am Acad Pol Soc Sci* 660(1):78–97. <https://doi.org/10.1177/0002716215576104>
- Royce E (2018) *Poverty and power: the problem of structural inequality*. Rowman & Littlefield, Washington, DC
- Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088):533–536
- Sahlins M (1995) *How “natives” think: about Captain Cook, for example*. University of Chicago Press, Chicago
- Salvatier J, Wiecki TV, Fonnesbeck C (2016) Probabilistic programming in python using PyMC3. *PeerJ Comput Sci* 2:e55
- Schneider A, Ingram H (1990) Behavioral assumptions of policy tools. *J Politics* 52(2):510–529. <https://doi.org/10.2307/2131904>

- Shultz KS, Whitney DJ, Zickar MJ (2013) *Measurement theory in action: case studies and exercises*, 2nd edn. Routledge, Oxfordshire
- Small ML (2009) ‘How many cases do i need?’: On science and the logic of case selection in field-based research. *Ethnography* 10(1):5–38
- Sriram A, Jun H, Satheesh S, Coates A (2017) “Cold fusion: training Seq2Seq models together with language models.” *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/1708.06426>.
- Star SL, Griesemer JR (1989) Institutional ecology, translations’ and boundary objects: amateurs and professionals in Berkeley’s museum of vertebrate zoology, 1907–39. *Soc Stud Sci* 19(3):387–420
- Tambe M, Adibi J, Al-Onaizan Y, Erdem A, Kaminka GA, Marsella SC, Muslea I (1999) Building agent teams using an explicit teamwork model and learning. *Artif Intell* 110(2):215–239
- Valiant LG (1984) A theory of the learnable. *Commun ACM* 27(11):1134–1142
- Wasserman S, Katherine F (1994) *Social network analysis methods and applications*. Cambridge University Press, Cambridge
- Weber M (1958) Science as a vocation. *Daedalus* 87(1):111–134
- Wilson WJ (1987) *The truly disadvantaged Chicago*. University of Chicago Press, Chicago
- Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1(1):67–82
- Wolpert DH, Macready WG (2005) Coevolutionary free lunches. *IEEE Trans Evol Comput* 9(6):721–735
- Wood F, Meent WF, and Mansinghka V. 2014. “A new approach to probabilistic programming inference.” In S Kaski & J Corander *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. PMLR(33):1024–1032.
- Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SYu (2020) A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst*. <https://doi.org/10.1109/TNNLS.2020.2978386>
- Xyrichis A, Ream E (2008) Teamwork: a concept analysis. *J Adv Nurs* 61(2):232–241
- Yung L, Louder E, Gallagher LA, Jones K, Wyborn C (2019) How methods for navigating uncertainty connect science and policy at the water-energy-food nexus. *Front Environ Sci*. <https://doi.org/10.3389/fenvs.2019.00037>
- Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, Wang L, Li C, and Sun M (2020) Graph neural networks: a review of methods and applications. *AI Open* (1):57–81. <https://doi.org/10.1016/j.aiopen.2021.01.001>

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Chris Graziul is a data scientist for the Urban Resiliency Initiative in the Department of Comparative Human Development at the University of Chicago. His research integrates diverse theoretical approaches and methodological techniques to better understand the mechanisms that structure social systems, especially the co-constitution of physical context and social behavior.

Alexander Belikov is a visiting scholar at Knowledge Lab at the University of Chicago (and the Head of Data Science at Hello Watt, Paris). In his research he merges computational techniques, such as network science, natural languages processing, neural networks and graphical models to identify universal patterns in social phenomena.

Ishanu Chattopadhyay is an Assistant Professor of Medicine at the University of Chicago, specializing in Machine Learning, AI and data science. Professor Chattopadhyay’s research focuses on the developing core algorithmic principles for modeling complex systems. As the Principle Investigator of several DARPA programs and the winner of the 2020 DARPA Young Faculty Award, Dr. Chattopadhyay’s interests span complex phenomena in biology, clinical decision-making, epidemiology, and social interactions. The Chattopadhyay laboratory focuses on the design of learning algorithms that push the limits of the current data science revolution. His work resides at the cusp of artificial intelligence, statistical theory,

formal languages, dynamical systems, and machine learning; aiming to formulate modeling approaches that work in the absence of subject matter experts, hopefully answering questions that we have not yet thought to ask.

Ziwen Chen is a MA student in the Computational Social Science program at the University of Chicago. Her research focuses on using large-scale data and advanced computational methods to study complex human behaviors, including urban mobility, digital culture, and business innovation.

Hongbo Fang is a PhD student in the Institute for Software Research at Carnegie Mellon University. His research focuses on the empirical study of online collaboration and social capital in the context of online interaction, and aims to understand the logic of internet-based voluntary participation and collaborative behaviors. Before joining CMU, he obtained his bachelor degree in computer science at Zhejiang University in 2019.

Anuraag Girdhar is an MA student in the Computational Social Science Program at the University of Chicago. His research uses discussion-based games to explore the emergent properties of social network structure, including political polarization, social theory of mind, and the wisdom of the crowd.

Xiaoshuang Jia is a Ph.D. student in the School of Sociology and Anthropology at Sun Yat-sen University. She was a visiting scholar at Knowledge Lab at the University of Chicago during December 2017—December 2018. In her research, she uses computational methods such as machine learning and social network analysis to understand social stratification and social structure. She also explores advanced techniques to do causal inference while taking the heterogeneity of population into account.

P. M. Krafft is a Senior Lecturer and MA Internet Equalities Course Leader at the University of the Arts London Creative Computing Institute. Dr. Krafft's research, teaching, and organizing aim to bridge computing, the social sciences, and public interest sector work towards the goals of social responsibility and social justice. Dr. Krafft pursues multiple programs towards this end. Much of Dr. Krafft's research centers around using mixed-methods social data science and participatory action research to study beliefs, ideology, and institutions in the information society, with a special focus on these topics in relation to artificial intelligence (AI) and other information systems. Dr. Krafft also conducts research in computer science and cognitive science with contributions to the areas of distributed AI, multiagent systems, human-computer interaction (HCI), and Bayesian modeling.

Max Kleiman-Weiner is Co-founder and CEO of Common Sense Machines. He was previously a Data Science Institute Fellow at Harvard and completed his Ph.D. in Brain and Cognitive Sciences at MIT in 2018. His research focuses on reverse-engineering the way the people build and use mental models of objects, places, and agents. He has focused on social aspects of human cognition and machine learning including cooperation, coordination, competition, social learning, and morality.

Candice Lewis is the Assistant Director of Knowledge Lab at the University of Chicago. She completed her Ph.D. in genetics from Penn State University in 2010. Dr. Lewis has worked on broadening participation and development of education programs at the University of Chicago for a decade.

Chen Liang is a MA student in the Computational Social Science Program at the University of Chicago. She graduated from the University of Michigan with a major in public policy. Her research focuses on using social network analysis and natural language processing techniques to understand the political polarization among policy experts in the United States.

John Muchovej is a Research Assistant in the Computation, Cognition, and Development Lab at Harvard, where he focuses on how people develop their common-sense understanding of the world. Previously, he was an undergraduate at the University of Central Florida and an RA at MIT. Past and current research focuses on using computational tools to better understand how we summarize extract intent from language and action observation and how people drastically narrow solution spaces in question-answering and creative thinking.


Alejandro Vientós is a PhD student at Rutgers-Newark working in the CoDaS Lab. Past and current work

centers around social inference, trust, and cooperation in multi-agent, incomplete information settings. Research interests include the design of games or mechanisms where cooperation is both a stable and optimal strategy and systems for incrementally converging to such strategies in said domains.

Meg Young is a postdoctoral fellow at Cornell Tech as part of the Digital Life Initiative in New York City. She uses ethnographic methods to study government technology. Much of her work explores artificial intelligence and public policy, with a focus on public–private partnerships, procurement, and accountability. Her other work explores artificial intelligence research and development through an organizational lens. She completed her PhD from University of Washington Information School.

James Evans is Max Palevsky Professor of Sociology, Director of Knowledge Lab, and Faculty Director of Computational Social Science at the University of Chicago and External Faculty at the Santa Fe Institute. He is Editor of the new *Journal of Social Computing* (IEEE). His research uses large-scale data, machine learning and generative models to understand how collectives think and what they know. This involves inquiry into the emergence of ideas, shared patterns of reasoning, and processes of attention, communication, agreement, and certainty. Thinking and knowing collectives like science, Wikipedia or the Web involve complex networks of diverse human and machine intelligences, collaborating and competing to achieve overlapping aims. Much of Evans' work has investigated modern science and technology to identify collective biases, generate new leads taking these into account, and imagine alternative discovery regimes. Evans also explores thinking and knowing in other domains ranging from political ideology to popular culture.

Authors and Affiliations

Chris Graziul¹ · Alexander Belikov¹ · Ishanu Chattopadhyay¹ · Ziwen Chen¹ · Hongbo Fang² · Anuraag Girdhar¹ · Xiaoshuang Jia³ · P. M. Krafft⁴ · Max Kleiman-Weiner^{5,6} · Candice Lewis¹ · Chen Liang¹ · John Muchovej^{5,6} · Alejandro Vientós^{5,7} · Meg Young⁸ · James Evans^{1,9} 

✉ James Evans
jevans@uchicago.ed

- ¹ University of Chicago, Chicago, USA
- ² Carnegie Mellon University, Pittsburgh, USA
- ³ Sun Yat-sen University, Guangzhou, China
- ⁴ University of Oxford, Oxford, England
- ⁵ MIT, Cambridge, USA
- ⁶ Harvard University, Cambridge, USA
- ⁷ Rutgers University, New Brunswick, USA
- ⁸ Cornell University, Ithaca, USA
- ⁹ Santa Fe Institute, Santa Fe, USA