Computational and Mathematical Organization Theory
https://doi.org/10.1007/s10588-021-09346-9

S.I. : GROUND TRUTH: IN SILICO SOCIAL SCIENCE (GTIS3)

The Ground Truth program: simulations as test beds for social science research methods

Asmeret Naugle[1] · Adam Russell[1] · Kiran Lakkaraju[1] · Laura Swiler[1] · Stephen Verzi[1] · Vicente Romero[1]

Accepted: 13 October 2021

Abstract
Social systems are uniquely complex and difficult to study, but understanding them is vital to solving the world's problems. The Ground Truth program developed a new way of testing the research methods that attempt to understand and leverage the Human Domain and its associated complexities. The program developed simulations of social systems as virtual world test beds. Not only were these simulations able to produce data on future states of the system under various circumstances and scenarios, but their causal ground truth was also explicitly known. Research teams studied these virtual worlds, facilitating deep validation of causal inference, prediction, and prescription methods. The Ground Truth program model provides a way to test and validate research methods to an extent previously impossible, and to study the intricacies and interactions of different components of research.

Keywords Human domain · Causal ground truth · Simulation test beds · Metascience · Complexity

1 Introduction

With growing appreciation for the national security challenges arising from *competitions short of war* (Jones 2019), *gray zone competitions* (Gray Zone Project 2020), *hybrid warfare* (Chivvis 2017), and operations that are *all-domain* (Underwood 2020), one domain is proving to be particularly difficult but particularly critical: the Human Domain (Gregg 2016; USSOCOM 2015; Bryant 2018; Branch 2021). As

✉ Asmeret Naugle
abier@sandia.gov

[1] Sandia National Laboratories, Albuquerque, NM, USA

Published online: 18 April 2022

an emerging term of art, the Human Domain invokes—but builds upon—the more common social science notion of social systems. The term is used here (and elsewhere[1]) to highlight the challenges of solving real world problems in the face of dynamic and evolving social complexities that we often barely understand. While every other domain—cyber, maritime, space, land, and air—presents its own significant challenges, the social systems that define the Human Domain present unique complexity, since they are defined and shaped by increasingly interconnected sociotechnical systems where strategic agents, collective fictions, emergent and mutable norms and rules, and changing motivations and values end up meeting and interacting (Russell 2019). Given this complexity, the study of social systems as manifest in the Human Domain presents a set of challenges beyond what is seen in other scientific disciplines. For those who believe that social science problems have simple solutions, it is worth considering the caveats of two Nobel laureate physicists: Murray Gell-Mann purportedly noted how much harder physics would be if electrons could think, and Richard Feynman opined how much harder physics would be if electrons had feelings (Lo 2010).

In spite of this complexity, researchers and decision makers still need reliable ways of understanding complex social systems and the dynamics that drive them. Since the social sciences have worked for centuries to find ways of studying and understanding the causal processes that shape system behaviors, and to predict and influence those systems, it would seem logical to turn to the social sciences to provide understanding, predictions, and recommendations for the Human Domain. However, while the social sciences have developed many important statistical, empirical, and computational methods for understanding these systems, studying social systems remains extremely difficult.

Reasons for this difficulty are abundant. For example, the causal processes that drive system behavior (or "causal ground truth") are almost never known or observable in sociotechnical systems; experimentation on real world human systems is extremely difficult due to ethical considerations as well as the inability to separate people from their environments or implement controlled experiments (Gerber and Green 2011); data, measures, and models are often biased (Olteanu et al. 2019); methods are designed for and applied to different social systems with different characteristics, limiting the comparability of those methods; and other concerns regarding the reproducibility and robustness of social science research methods and models, stemming in part from evidently pervasive "Questionable Research Practices"

[1] For more information on the origin and use of the term "Human Domain," its definitional and conceptual challenges, and why it is used in place of more common social science terms like "social systems", please see https://smallwarsjournal.com/jrnl/art/does-the-human-domain-matter; https://smallwarsjournal.com/jrnl/art/joint-force-2020-and-the-human-domain-time-for-a-new-conceptual-framework; https://www.soc.mil/SWCS/SWmag/archive/SW2703/human%20domain.pdf; https://othjournal.com/2019/06/17/an-approachable-look-at-the-human-domain-and-why-we-should-care/; https://mwi.usma.edu/the-pentagon-bureaucracy-and-the-human-domain-of-war/; https://apps.dtic.mil/dtic/tr/fulltext/u2/a623748.pdf; https://smallwarsjournal.com/jrnl/art/operationalizing-science-human-domain-great-power-competition-special-operations-forces; https://nsiteam.com/social/wp-content/uploads/2017/01/SOF-OHD-Concept-V1.0-3-Aug-15.pdf; https://www.army.mil/article/141535/future_joint_concepts_focus_on_human_elements.

(Munafò et al. 2017), as well as a general dearth of efforts (and funding) to replicate or evaluate the robustness of certain methods across different systems, contexts, and under different conditions. Further, there are increasingly sophisticated, data-driven efforts that are being adopted by the social sciences (Zhang et al. 2020), which increasingly depend upon machine-learning tools that enable researchers to model human systems using previously-impossible numbers of parameters, but without any real sense of whether those methods are better than more traditional social science approaches, and no seeming way to quantify the improvements if they are. Finally, future social science research is likely to be conducted by increasingly multidisciplinary teams (Lazer et al. 2020), but in the absence of having ways to best understand how and why certain disciplines and skill sets might be used for tackling different kinds of social systems challenges, teaming may remain unnecessarily ad hoc.

To try to make progress in the face of these seemingly intractable problems for enhancing social science capabilities, DARPA's Ground Truth program was devised as a new way of testing a range of social science methods. The program addressed common difficulties by shifting away from testing methods on real-world systems, and instead testing the methods on carefully designed and controlled simulations of social systems. The simulations had known causal ground truth, known future states, and were able to produce counterfactual data, all key features that allowed explicit validation of social science research methods in ways completely unavailable from real-world systems. Data availability was controllable, as was the complexity of the simulation test beds. Researchers were able to collect their own data from the simulations, including conducting various kinds of experiments, which led to far more targeted, and potentially less biased, data availability than would typically be available from a real-world system. By using the same simulations as test beds for a variety of research methods, and by ensuring there was causal ground truth and data needed for their explicit validation, the program also allowed true comparability of methods across a range of metrics.

While there were multiple potential payoffs for the high-risk work that Ground Truth undertook, two key motivating questions guided the design of the program. First, are social simulations useful test beds for social science research methods? The program developed and deployed multiple simulations to compare their utility and tease out characteristics of the simulations that may have made them useful as test beds. Second, what could the program teach us about the abilities and limitations of a range of social science and data-driven research methods? With multiple research teams and multiple test beds, the program was able to compare explicitly validated results, analyze trends, and determine which methods seemed to work better for studying different types of social systems.

2 Program structure

The Ground Truth program was organized around three challenges, each consisting of three tests: explain, predict and prescribe (Table 1). The complexity of the social simulation test beds varied over the challenges, reflecting a range of types of complexity (Castellani 2014), while each type of test assessed a different set of

Table 1 Program challenges, tests, and timeline

Initial development	Challenge 1			Challenge 2			Challenge 3		
9 Months	9 Months			6 Months			6 Months		
	Explain test	Predict test	Prescribe test	Explain test	Predict test	Prescribe test	Explain test	Predict test	Prescribe test
Simulations and research approaches developed	3 Months	3 Months	3 Months	2 Months	2 Months	2 Months	2 Months	2 Months	2 Months

capabilities and a different ambition of social science. The explain tests asked the research teams to determine the *causal structure* of the system and tested their causal inference abilities. What influenced the actors, and what determined the behaviors of both the people and their environments? What were the important variables, and how were they causally connected to each other? The predict tests involved specific questions for the research teams to answer. The predict questions were specific to each social simulation test bed, but included near and far term predictions, baseline and counterfactual ("what if") predictions, and assessments of uncertainty. Prescribe tests examined the research teams' ability to find ways of influencing the test bed systems to achieve desired goals. The prescribe tests involved simulation-specific question sets, which included questions covering multiple levels of analysis at the individual, group, and system levels.

Three categories of teams participated in the program, including the social simulation teams, the research teams, and a test and evaluation team (which for this discussion includes program design/management and other government partners). Four simulation teams designed and built the simulation test beds (Parunak 2022; Pynadath et al. 2022; Rager et al. 2022; Züfle et al. 2022). The simulations focused on different Human Domain topics and systems (urban life, financial governance, disaster response, and geopolitical conflict), but all of the simulations involved actors making decisions, implementing behaviors, and interacting with each other and their environments. Accordingly, the simulations were considered *virtual worlds*. The simulations were not meant to reproduce real world systems, and the simulation teams were discouraged from describing them as such, although the teams were asked to design social simulations that would be plausible as virtual worlds (for example, in no world did agents have total omniscience or reproduce via parthenogenesis). Each simulation was thoroughly evaluated by the test and evaluation team using a range of measures described in more detail below, either before each challenge launched or early in the challenge timeline. The simulation teams also developed, in consultation with the test and evaluation team, the specific questions and details for each test. Key research challenges presented by each simulation, as well as the relative importance of agency versus structure to each, is presented in Fig. 1.

To begin each test, each simulation team produced an initial data package for the research teams. The initial data packages included carefully selected sets of data collected from the simulations, as well text documents with any information the research teams would need. The research teams were then allowed to collect data from the simulations, using a constrained but representative set of research and experimental approaches and methods (see "accessibility to research methods" section for more detail). The simulation and research teams were entirely firewalled, and all communication between them was conducted through, and filtered by, the test and evaluation team. This ensured that team identities were kept unknown (to the greatest extent possible), and that no information would pass between teams that did not fit with the program concept. For example, the simulation teams were not allowed to directly release information during the program about the causal ground truth driving the simulations (although this did not prevent the research teams from trying, which justified the firewall principle).

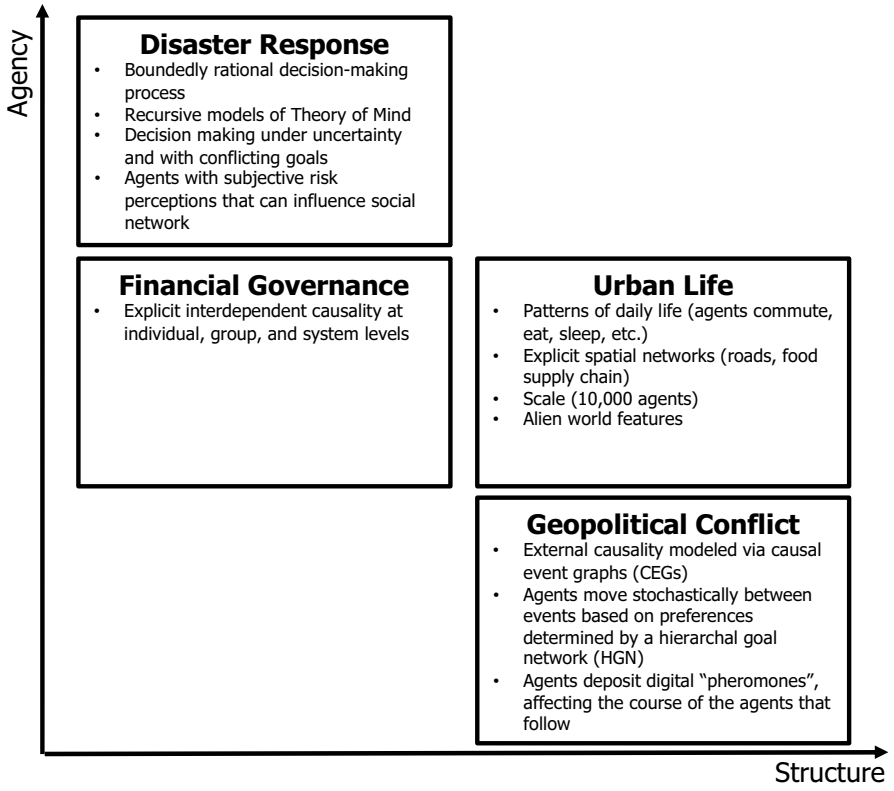


Fig. 1 Key research challenges and relative importance of agency and structure for the four virtual worlds

To collect data, the research teams submitted research requests to the test and evaluation team, which were evaluated, logged, and passed on to the simulation teams. A question and answer process was also enacted, so that the research teams and simulation teams could communicate about what types of research the simulations enabled, what research questions made sense within the context of the simulations, and how information passed between the teams should be interpreted. The simulation teams then implemented the research requests, produced the applicable data from the simulations, and sent the data to the test and evaluation team, who evaluated the data before sending it on to the research teams.

At the end of each test, the research teams submitted results to the test and evaluation team. More information on the methods used to develop these submissions can be found in Graziul et al. (2022), Schmidt et al. (2022), and Volkova et al. (2022). The test and evaluation team collected any necessary data (for example, historical projects from the simulations) and evaluated the research teams' performance. Evaluations were shared with the research teams, so that their performance on each test was known early in data collection process of the following test. The evaluations did not include specific feedback (such as the subsets of the

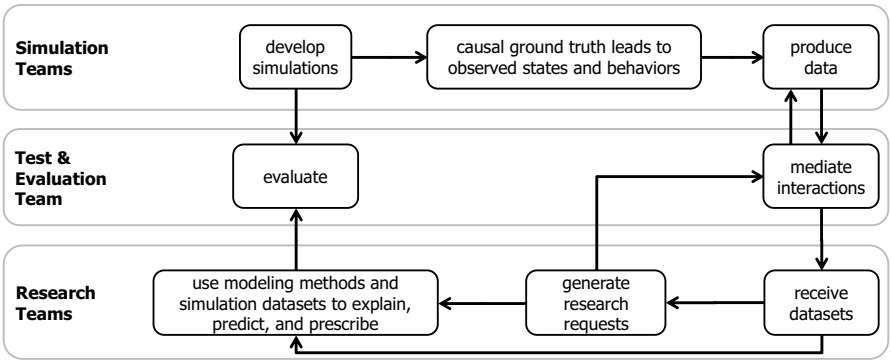


Fig. 2 Key roles and interactions between teams

inferred causal ground truths that were correct), instead focusing at a higher level (such as the fraction of elements in the inferred ground truths that were correct). An overview of the key roles and interactions between the simulation, research and test and evaluation teams is shown in Fig. 2.

Three research teams participated in the program. Two of the teams participated in all nine tests (explain, predict, and prescribe over each of three challenges). The final research team had a different role; they did not collect data, instead working with the datasets collected by the other teams, as well as full datasets, for Challenge 1 of the program. The purpose of this team was to study the same systems using different datasets, to provide insight into the relative importance of data as compared to methodology.

3 Evaluating the simulations

Before each program challenge, the test and evaluation team evaluated the simulations. Evaluation metrics were chosen around two goals. First, we wanted to ensure that the simulations would be useful as test beds for the research methods being used. The metrics that were evaluated in this category included the accessibility, verifiability, flexibility, and plausibility of the simulations. The second goal of simulation evaluation, which focused on simulation complexity, was to better understand the situations in which the research teams did well (or poorly), and potential contributing factors.

Some of the simulation evaluation metrics, as well as some of the evaluation methods for the research teams, relied on explicit graphical representations of the virtual worlds' causal structures. We called these representations the *causal ground truth graphs*, and specified guidelines for their representation of causality. Each node in a causal ground truth graph represents a variable in the associated simulation. Nodes are connected by directed edges representing causal relationships, so that an edge going from node A to node B would indicate that variable A was used in the equation/algorithm that determined the value of variable B.

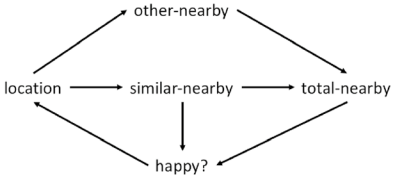
Pseudo Code (adapted from Wilensky 1997)	Causal Ground Truth
create turtles on random patches in a grid at each time step: if all turtles are happy then stop for each turtle if unhappy, randomly move to new unoccupied patch similar-nearby count = number of neighbors with color = turtle's color other-nearby count = number of neighbors with color != turtle's color total-nearby = similar-nearby + other-nearby happy? = yes *if* similar-nearby >= (%-similar-wanted * total-nearby/100)	

Fig. 3 Example of a causal ground truth graph, for the Schelling segregation model

The test and evaluation team provided guidance on how the ground truth graphs should be developed, so that the simulation and research teams were applying the same processes. These guidelines included aggregating nodes and edges where possible, and not including parameters (which had no capacity to change during the simulation time horizon) or initial values as nodes. To illustrate the concept of a causal ground truth graph, Fig. 3 shows such a graph for the well-known Schelling segregation model (Schelling 1971) as implemented in the in the NetLogo (Wilesnky 1999) models library (Wilensky 1997). The simulation virtual worlds used for the Ground Truth program were significantly more causally complex than this example, but the causal ground truth graphs were developed using the same principles.

3.1 Accessibility to research methods

The first evaluation metric for the Ground Truth simulations was simulation accessibility, or the simulation's ability to accommodate the range of data collection methods that might be used by social science research teams. To push the state of the art in this area, research teams were encouraged to use a wide range of methods and were asked to provide lists of research methods that they might want to apply to the simulations. These methods ranged from highly qualitative methods (such as interviews) to highly quantitative methods (such as regression and time-series analyses, as well as more advanced machine-learning approaches). To ensure that the research teams were able to use a sufficient range of methods, the simulation teams needed to be able to interpret requests, conduct analyses, and return the data appropriate to the methods. To test accessibility, the test and evaluation team developed accessibility demonstrators, simple examples of research requests for each method that were targeted toward the respective simulations. The simulation teams answered these accessibility demonstrators as they would research requests and were evaluated based on the percent of methods to which they were accessible. The simulation teams were asked to strive to be accessible to 100% of the proposed methods and were required to be accessible to at least 50% in order to continue in the program.

3.2 Verifiability of causal ground truth

To achieve the goals of the Ground Truth program, a key requirement for the social simulations was the ability to define the exact causal processes that drive the simulation's agent, group, and system behavior. In the real world we can try to derive causal inferences using data, but simulation virtual worlds allow for explicit validation of causal inference by comparison of inferred and true causal ground truths. The Ground Truth program evaluated simulations on their verifiability to assure that the simulations had this quality. Along with the metrics discussed below, the simulation teams were required to do unit testing and verification of their simulations.

The set of metrics for evaluating verifiability tested the similarity between the simulation's actual causal structure and the causal ground truth graph representation of that structure. The simulation teams submitted their ground truth graphs and annotated simulation code for the verifiability evaluations. The test and evaluation team identified and evaluated two metrics based on these: completeness and precision. The completeness metric measured the fraction of variables (nodes) and relationships between variables (edges) in the simulation equations that were included in the ground truth graph. Precision measured the fraction of ground truth graph edges and nodes that were included in the simulation equations. The simulation teams were asked to strive for 100% completeness and precision to ensure that their ground truth graphs would accurately reflect their simulations, and thus serve as appropriate grading rubrics for the research teams' causal inference efforts.

3.3 Complexity

The Ground Truth simulations were meant to serve as proxies for the real world. The simulations needed to be simple enough to provide reasonable tests for the research teams, yet complex enough to emulate characteristics that research methods might encounter in real world situations. The original program concept was that simulation complexity should increase through successive challenges, from simple systems in Challenge 1 to organized complex systems in Challenge 3. Complexity was also considered a potential indicator of the difficulty of understanding the simulations, useful for comparing the performance of the research teams across simulations and challenges.

While many definitions of complexity have been proposed (Ladyman 2013; Mitchell and Newman 2002; Bar-Yam 2002; Watts 2017), no definition is widely accepted. To ensure that an appropriate range of characteristics were considered, the test and evaluation team identified a suite of metrics that captured different elements of simulation complexity. Using a carefully chosen combination of methods, we hoped to gain a deeper and more nuanced understanding of simulation complexity than could be achieved with a single metric.

Complexity metrics were chosen based on an organizing structure involving two dimensions applicable to the Ground Truth program (Table 2). The first dimension differentiated between metrics that require knowledge of the causal

Table 2 Organizing structure for choosing complexity metrics

	Not tied to social/behavioral sciences	Inspired by social/behavioral sciences
Requires knowledge of system structure	Causal complexity	Number of differentiated relationships
Requires no knowledge of system structure	Forecast complexity	Global reaching centrality

structure (or ground truth graph) of the simulation and those that do not. Metrics that require knowledge of causal structure are useful for causal simulations, but we generally do not have knowledge of the causal structure of real-world systems; we thus chose metrics covering each of these categories. The second dimension related to the application space, distinguishing between metrics that are inspired by the social and behavioral sciences and those that are more broadly applicable.

The test and evaluation team developed a metric of causal complexity, which required knowledge of system structure and was not tied to the social and behavioral sciences. This metric integrated cyclomatic complexity (which captures the interconnectedness of a graph) and feedback density (the fraction of the ground truth diagram's nodes and edges that are involved in at least one feedback loop) (Naugle 2019). Causal complexity captured how complicated and intricate the causal structure, represented by the causal ground truth graph, was for each simulation.

Approximate forecast complexity, which measures the information content of a simulation's output data, was selected as the metric that did not require knowledge of the system structure and was not tied to the social and behavioral sciences. Forecast complexity (Shalizi 2006) is the amount of information needed for optimal prediction where a past portion of a time series is used to predict a future portion. The test and evaluation team approximated forecast complexity by using Normalized Compression Distance (*NCD*) (Cilibrasi and Vitanyi 2005), an estimate of Normalized Information Distance (*NID*), between past and future data. Normalized compression distance captures the similarity between two data sets.

The number of differentiated relationships was selected, because of its use in real-world social science research, as the metric that required knowledge of the system structure and was tied to the social and behavioral sciences. For the purposes of the Ground Truth program, this metric was calculated as the number of distinct types of relationships available to actors in a simulation, as defined in the causal ground truth.

The final complexity metric, global reaching centrality (Mones 2012), requires no knowledge of the system structure and is inspired by the social and behavioral sciences. This metric quantifies hierarchy in a social network by considering the distribution of reach centralities, or the number of nodes that can be reached within a defined number of steps, in that network. Global reaching centrality was calculated on the social network resulting from a simulation, and was interpreted as a measure of the emergent organization of that social network.

While there were no targets or benchmarks for these metrics, simulation complexity was used to compare the simulations across challenges and simulation teams.

3.4 Flexibility

To ensure that the simulations could systematically control and increase their complexity over the successive challenges of the Ground Truth program, they were evaluated on their flexibility. The simulation teams were asked to ensure that at least 30% of their simulation parameters had potential to significantly increase or decrease simulation complexity. This was tested using a main effects analysis (Vik 2013) that compared the forecast complexity and global reaching centrality for a baseline result versus parameter-varied simulation scenarios. To implement this analysis, simulation teams generated sets of runs with varied input parameters, and complexity metrics were calculated for each of those runs. Flexibility was evaluated by testing for significant differences among those complexity metrics, indicating whether the varied input parameters had a significant impact on the complexity of the simulated world. The range (upper and lower bounds) of the potential forecast complexity and global reaching centrality were also measured to indicate the controllability of the simulation's complexity.

3.5 Plausibility

Plausibility metrics were designed to test each simulation's credibility as a virtual world test bed. Plausibility evaluations tested the simulations' social plausibility as well as their ability to provide non-trivial results without requiring external intervention. Social plausibility was tested based on qualitative requirements that increased in realism as program progressed: in Challenge 1 the simulations were required to include multiple actors interacting with each other, in Challenge 2 those actors were required to form groups that interacted both with actors and with other groups, and in Challenge 3 those actors and groups were required to influence the system or environment in which they resided. To establish non-triviality of results, the simulation teams were required to show that the normalized entropy of their simulation results was not equal to minimum or maximum entropy (which would have indicated unchanging or random behavior respectively), and that the variance of quantities of interest over time was not zero (which would have indicated unchanging quantities of interest).

4 Evaluating the research methods

The research teams were given three different types of tests: explain, predict, and prescribe. The test and evaluation team evaluated their performance for accuracy, robustness, and efficiency.

4.1 Explain test accuracy

Validating causal structures is extremely difficult in real-world situations. Validation generally focuses on assessing the similarity of predictions to observed behaviors, which does not account for causality because different causal mechanisms might generate similar results. In the physical modeling field, this is known as non-uniqueness or non-identifiability of explanatory or causal structure (Oreskes 1994). Because of this, methods for validating a model's causal structures against the real world are limited. Since the Ground Truth simulations were built around explicit, known causal ground truths, the program facilitated explicit causal validation of the research teams' results.

The research teams' explain test goal was to infer the causality of each simulation. To assess the accuracy of these inferences, the test and evaluation team compared the research teams' inferred ground truth graphs to the ground truth graphs representing each simulation's causal structure. Research teams submitted inferred ground truth graphs, and the simulation teams mapped the nodes (or variables) in those graphs to nodes in the simulation ground truth graphs. This allowed for translation between the graph structures and facilitated quantitative comparison. The inferred and simulation graphs were then assessed for precision (the fraction of inferred items that were in fact in the simulation ground truths) and recall (the fraction of items from the simulation ground truths that were inferred) for both edges and nodes, and F1 scores (the harmonic mean of precision and recall) (Rijsbergen 1979; Yang 1999).

4.2 Predict test accuracy

The predict tests focused on a common goal of data analysis and science in general: forecasting future states of the system of interest. The use of simulation test beds allowed not only for single-state prediction to be validated, but also for explicit generation, exploration, and comparison of counterfactuals and alternative scenarios. The predict tests focused on specific questions generated by the simulation teams, and appropriate evaluation metrics were identified for each question. In general, the evaluations focused on metrics such as differences between point estimates and Jaccard indices for comparing sets.

4.3 Prescribe test accuracy

The prescribe tests examined the research teams' ability to prescribe system interventions to influence the simulations in desired ways. As with the predict tests, prescribe tests varied, and were defined by question sets identified by the simulation teams. In general, prescriptions were made by the research teams, and the simulation teams implemented those prescriptions in their simulations to determine their results. Those results were then compared to a baseline (with no prescriptions implemented), as well as to optimal or best available results, based on analysis by the

simulation teams. Prescribe results were evaluated based on the specific questions posed and included things like the percentages of prescribed actions that produced responses in a desired direction (increasing/decreasing) as compared to a baseline, and percentage of the distance between baseline and target outcome achieved by the prescription.

4.4 Robustness

Robustness refers to how well a research method performed over a range of applications of the method. Robustness was measured as an indication of whether a method is useful for a range of applications. This is similar to the concept of generalizability. For example, a social science modeling method might be very good at predicting the results of a very specific scenario for which it was developed. If that method doesn't do well at predicting results of other scenarios, then it will be considered less robust than a method that performs well over many scenarios. To evaluate robustness, this program measured average accuracy, both over different questions from a single simulation, and over multiple simulations.

4.5 Efficiency

Many different methods could be used to explain, predict, and prescribe social systems. The efficiency of those methods, including how long it takes a TA2 team to apply the methods, how much computational power is required, and how much data is required, affects the functional utility of the methods. More efficient methods should require less effort and may therefore be more applicable in real-world situations. For the purposes of this program, we focused efficiency evaluations on how much data was attained by the research teams to produce results. Efficiency was calculated as the bytes of data, after compression, received by a research team from a virtual world simulation.

5 Program evolution

The Ground Truth program was designed to use social simulations as plausibly realistic proxies for real world social systems, with complexity of the simulation test beds increasing over the course of the program. Some of the aspects of the Challenge 1 tests that were meant to reflect real-world complexity proved to be overly difficult for the research teams. For example, the simulation teams were instructed in Challenge 1 to limit data collection to levels that would seem reasonable in the real world. Initial data packages were small, as is common when approaching a new problem or social system, and the simulation teams implemented limits on what percentage of their populations could be sampled using particular research methods (in reality it is not plausible to survey every human being, all the time, about anything). This presented major frustrations for the research teams, who realized that part of operating in the Human Domain means having to become acquainted with the

"worlds" in question, determining what types of research could and should be done on each simulation, and collecting sufficient data to implement analytic methods.

To adjust for this difficulty, Challenge 2 of the program shifted in character from the original program plan. The simulation teams were asked not to make their simulations more complex in Challenge 2, instead keeping very similar causal ground truths as in Challenge 1 and maintaining or even reducing complexity. The simulation teams also relaxed some of the realism-focused restrictions around data collection, allowing substantially higher sampling rates as compared to Challenge 1.

In Challenge 3, simulation complexity increased as originally planned. The research teams indicated that data availability was still a major bottleneck in their performance, thus driving a decision to increase data availability substantially for Challenge 3, making what would be in many cases an ideal scenario for social scientists. For the explain test, the simulation teams were encouraged to include quite a lot of data in their initial data packages, and to allow very high data collection rates through the research process. For the predict and prescribe tests, the research teams were given full datasets from the simulations, including all observable data produced over the observed timelines. The teams were still allowed to conduct experiments to collect further data. After the initial Challenge 3 explain test, the research teams were given access to full datasets from Challenges 1 and 3 to reconduct the explain tests.

6 Conclusions

The Ground Truth program showed that simulations can be used as test beds for research methods in the Human Domain, serving as a proof of concept of a new way for social science to build upon its current paradigm of "discover and know" towards a paradigm of "create and solve" (Prabhakar 2020). The Ground Truth simulations served as functional test beds for data collection methods, and provided abundant data and information on causal structures, potential future states, and the results of policy prescriptions and system manipulations. This abundance of data – produced under conditions of controlled complexity and known ground truth—allowed for *explicit* validation, not only of straightforward predictions, as is typically done for validation exercises of predictions against real-world data, but also of causal inference results and counterfactuals.

The explain, predict, and prescribe tests gave the program broad applicability to real world problems. Explaining the causal structure of a social system can help a decision maker to understand the underlying causes and solutions to problems of interest, and has potential to improve both prediction and prescription. Prediction allows decision makers to anticipate and prepare for the future, and to understand likely impacts of our actions. Prescription is perhaps the true aspirational endeavor – to produce actions and interventions that achieve desired outcomes. The Ground Truth program brought these three real-world goals together, pushing the research teams to consider different key purposes of social science in context and in a controlled time frame.

The Ground Truth program design also offered a new opportunity for rich study of metascience. Research teams were tasked with end-to-end study of the simulations as virtual worlds. They collected data, ran experiments, analyzed information, and produced theories about how the virtual worlds worked. Each of these tasks is extremely challenging, and in the real world these activities are often siloed, with different researchers attacking each problem, their methods relying on the quality of previous work. While silos provide an opportunity to develop highly technical and intricate theories and algorithms, they reduce understanding of context and interaction. They also conceal how assumptions made in one silo affect other parts of the system. By tasking the research teams with a more holistic mission, this program design forced research teams to consider interactions among the typical silos. Abundant data was collected on data collection strategies and effectiveness, methodology, the evolution of approaches, and even the frustrations of the different teams. This offers an incredible opportunity for studying not only how social science is done, but also how the typical silos interact with each other, and opportunities for improvement of the field – that is to say, how social science might be done better. Note that we provide a detailed analysis of the Ground Truth results, including team performance and comparative metrics, later in this special issue (Naugle et al. 2022).

This removal of silos also makes a Ground Truth-like program valuable as a potential teaching tool. By participating in a controlled, end-to-end research task, with broad goals covering explain, predict, and prescribe tasks, researchers could begin to truly understand the breadth of methods, potential pitfalls, dependencies, and metascience in which their fields exist. This type of learning opportunity could also be vital for decision makers who need to be able to interpret social science findings in their full context.

This program came at a vital time in the evolution of social science as a field. As computational social science rises in prominence (Lazer 2020), there has been a growing demand from academia, industry, and government to better understand how the results of data-driven algorithms are influenced by assumptions, biases in data, and environmental conditions. As data-driven algorithms, spearheaded by the surge in artificial intelligence and machine learning capabilities, take more prominent roles in government and other decision making, it is even more important to understand these issues.

While the Ground Truth program successfully showed that simulations can be used as test beds for social science research methods, continued work in this area is needed to further test the concept and improve effectiveness. First, this program was designed to be plausibly representative of some real-world systems, with all of their associated complexity. The program incorporated four relatively complex simulations and allowed the research teams to tackle the associated tests in any ways they found useful. A similar program with a carefully controlled research design could further enhance our ability to test methods and understand their design utility. In such a program, simulations would progress from very simple (only a few variables and causal relationships) to more complex in a very structured manner, and research teams would apply exactly the same methods to the same data produced from the respective simulations. This would allow more unambiguous evaluation of the utility of specific social science research methods. Recent large-scale data analysis

programs (Salganik et al. 2020; Botvinik-Nezer et al. 2020) might serve as models, but have not utilized simulations with known causal ground truth and known complexity, thus lacking the capability for deep, explicit validation that the Ground Truth program provided.

Another major need is investigation into the generalizability of Ground Truth results to the real world. While it was out of scope for the initial program given all the other high-risk challenges that had to be first addressed, there remains the obvious question of how to assess whether—and to what degree—a research method that performs well on a simulation test bed would perform equally as well on an associated real-world system. One way to build on this program's achievements would be to study characteristics, including complexity, of simulation test beds and associated real-world systems that might indicate applicability of results. This study could incorporate a broad set of characteristics and metrics along with three validation efforts: one comparing the simulation test beds to the real-world systems of interest, one comparing research method results to those test beds, and one comparing the same research method results to the associated real-world systems.

The Human Domain is hard, which makes the social sciences – often referred to as the soft sciences – potentially the "hardest sciences" (White 2012). The Ground Truth program introduced a novel, controllable method for developing, testing, and validating a growing diversity of social science research methods. The study of social simulation test beds does differ from the study of social science; human systems under study in social science research are presumably much more complex and unpredictable than those in the simulations used for this program. More investigation is needed to understand exactly how simulation test beds should be used, but we believe this program showed that it is not only possible, but potentially enlightening to test social science research methods on social simulations with known ground truth.

Acknowledgements Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This project is sponsored by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreement no. HR0011937661. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy, the Department of Defense, or the United States Government.

Funding This study was funded by Defense Advanced Research Projects Agency (HR0011937661).

References

Bar-Yam Y (2002) General features of complex systems. Encyclopedia of life support systems. UNESCO, Oxford

Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, Kirchler M, Iwanir R, Mumford JA, Adcock RA, Avesani P (2020) Variability in the analysis of a single neuroimaging dataset by many teams. Nature 20:1–7

Branch A, Cardon E, Ellis D, Russell A (2021) We ignore the human domain at our own peril. Modern war institute. https://mwi.usma.edu/we-ignore-the-human-domain-at-our-own-peril/. Accessed 20 July 2021

Bryant S, Cleveland C, Jensen B, Arnel D (2018) Military strategy and the 21st century. Cambria Press, Amherst

Castellani B (2014) FOCUS: complexity and the failure of quantitative social science. Discover society. https://discoversociety.org/2014/11/04/focus-complexity-and-the-failure-of-quantitative-social-science/. Accessed 20 July 2021

Chivvis CS (2017) Understanding Russian "hybrid warfare" and what can be done about it. RAND Corporation, California

Cilibrasi R, Vitanyi PMB (2005) Clustering by compression. IEEE Trans Inf Theory 51:1523–1545

Gerber AS, Green DP (2011) Field experiments and natural experiments. The Oxford handbook of political science. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199604456.001.0001/oxfordhb-9780199604456-e-050. Accessed 20 July 2021

Gray Zone Project (2020) Center for Strategic and International Studies. https://www.csis.org/grayzone. Accessed 20 July 2021

Graziul C, Belikov A, Chattopadyay I, Chen A, Fang H, Girdhar A, Jia X, Krafft P, Kleiman-Weiner M, Lewis C, Liang C, Muchovej J, Vientós A, Evans J (2022) Does big data serve policy? Not without context. An experiment with in silico social science. Comput Math Organ Theory, this issue

Gregg H (2016) Human domain and influence operations in the 21st century. Spec Oper J 2(2):92–105

Jones RC (2019) Deterring "Competition short of war": are gray zones the ardennes of our modern maginot line of traditional deterrence? Small wars journal. https://smallwarsjournal.com/jrnl/art/deterring-competition-short-war-are-gray-zones-ardennes-our-modern-maginot-line. Accessed 20 July 2021

Ladyman J, Lambert J, Wiesner K (2013) What is a complex system? Eur J Philos Sci 3:33–67

Lazer DMJ, Pentland A, Watts DJ, Aral S, Athey S, Contractor N, Freelon N, Gonzalez-Bailon S, King G, Margetts H, Nelson A, Salganik M, Strohmaier M, Vespignani A, Wagner C (2020) Computational social science: obstacles and opportunities. Science 369:1060–1062

Lo A, Mueller M (2010) Warning: physics envy may be hazardous to your wealth! J Invest Manage 8(2):13–63

Mitchell M, Newman M (2002) Complex systems theory and evolution. Encyclopedia of evolution. Oxford University Press, Oxford

Mones E, Vicsek L, Vicsek T (2012) Hierarchy measure for complex networks. PLoS ONE 7:e33799

Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers EJ, Ware JJ, Ioannidis JPA (2017) A manifesto for reproducible science. Nat Hum Behav 1:1–9

Naugle AB, Swiler LP, Lakkaraju K, Verzi S, Warrender C, Romero V (2019) Graph-based similarity metrics for comparing simulations and causal loop diagrams. Sandia National Laboratories, Albuquerque

Naugle AB, Krofcheck D, Warrender C, Lakkaraju K, Swiler L, Verzi S, Emery B, Murdock J, Bernard M, Romero V (2022) What can simulation test beds teach us about social science? Results of the ground truth program. Comput Math Organ Theory, this issue

Olteanu A, Castillo C, Diaz F, Kıcıman E (2019) Social data: biases, methodological pitfalls, and ethical boundaries. Front Big Data 2:13

Oreskes N, Shrader-Frechette K, Belitz K (1994) Verification, validation, and confirmation of numerical models in the earth sciences. Science 263:641–646

Parunak V (2022) SCAMP's Stigmergic model of social conflict. Comput Math Organ Theory, this issue

Prabhakar A (2020) In the realm of the barely feasible. Issues in science and technology 37(1): 34–40

Pynadath D, Dilkina B, Jeong D, John R, Marsella S, Merchant C, Miller L, Read S (2022) Disaster world: decision-theoretic agents for simulating population responses to hurricanes. Comput Math Organ Theory, this issue

Rager S, Leung A, Pinegar S, Mangels J, Poole M, Contractor N (2022) Groups, governance, and greed: the ACCESS world model. Comput Math Organ Theory, this issue

Rijsbergen CJV (1979) Information retrieval, 2nd edn. Butterworth-Heinemann, USA

Russell A (2019) Fomenting (reproducible) revolutions: DARPA, replication, and high-risk, high-payoff research (Video). In: Metascience 2019 Symposium. https://www.metascience2019.org/presentations/adam-russell/. Accessed 20 July 2021

Salganik MJ, Lundberg I, Kindel AT, Ahearn CE, Al-Ghoneim K, Almaatouq A, Altschul DM, Brand JE, Carnegie NB, Compton RJ, Datta D (2020) Measuring the predictability of life outcomes with a scientific mass collaboration. Proceedings of the National Academy of Sciences. 117(15):8398–8403. https://www.pnas.org/content/117/15/8398. Accessed 20 July 2021

Schelling TC (1971) Dynamic models of segregation. J Math Sociol 1(2):143–186

Schmidt A, Cameron C, Lowman C, Brulé J, Deshpande A, Faemi S, Barash V, Greenberg A, Costello C, Sherman E, Bhattacharya R, McQuillan L, Perrone A, Kouskoulas Y, Fink C, Zhang J, Shpitser I, Macy M (2022) Searching for explanations: Testing social scientific methods in synthetic ground-truthed worlds. Comput Math Organ Theory, this issue

Shalizi CR (2006) Methods and techniques of complex systems science: an overview. In: Deisboeck TS, Kresh JY (eds) Complex systems science in biomedicine. Springer, US, Boston, pp 33–114

Underwood K (2020) The army shapes joint all-domain operations. SIGNAL Magazine. https://www.afcea.org/content/army-shapes-joint-all-domain-operations. Accessed 20 July 2021

United States Special Operations Command (2015) Operating in the human domain. Version 1.0. August 2015. https://nsiteam.com/operating-in-the-human-domain-1-0

Vik P (2013) Regression, ANOVA, and the general linear model: a statistics primer. SAGE Publications, California

Volkova S, Arendt D, Saldanha E, Glenski M, Ayton E, Cottam J, Aksoy S, Jeffereson B, Shrivaram K (2022) Causal discovery and prediction of human behavior and social dynamics from observational data: Generalizability, reproducibility and robustness. Comput Math Organ Theory, this issue

Watts DJ (2017) Should social science be more solution-oriented? Nat Hum Behav 1:1–5

White MD (2012) Which is "harder": social science or physical science? In: Economics and ethics. https://www.economicsandethics.org/2012/07/which-is-harder-social-science-or-physical-science.html. Accessed 20 July 2021

Wilensky U (1997) NetLogo Segregation model. http://ccl.northwestern.edu/netlogo/models/Segregation. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Wilensky U (1999) NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Yang Y (1999) An evaluation of statistical approaches to text categorization. Inf Retrieval 1:69–90

Zhang J, Wang W, Xia F et al (2020) Data-driven computational social science: a survey. Big Data Res 21:100145

Züfle A, Wenk C, Pfoser D, Crooks A, Kavak H, Kim J, Jin H (2022) Urban life: a model of people and places. Comput Math Organ Theory, this issue

Asmeret Naugle PhD is a system dynamics modeler with expertise in interdisciplinary models, hybrid modeling, and model assessment techniques. She focuses on computational social science research and application, including verification and validation (V&V), uncertainty quantification, and sensitivity analysis methods, with emphasis on developing methods for assessment of social-behavioral and complex systems models.

Adam Russell PhD is interested in new experimental platforms and tools to facilitate discovery, quantification, and "big validation" of fundamental measures in social science, behavioral science, and human performance. Dr. Russell has broad technical and management experience across a number of disciplines, ranging from cognitive neuroscience and physiology to cultural psychology and social anthropology. Dr. Russell holds a Bachelor of Arts in Cultural Anthropology from Duke University, and an M.Phil. and a D.Phil. in Social Anthropology from Oxford University, where he was a Rhodes Scholar.

Kiran Lakkaraju PhD has more than 10 years of experience developing models of human behavior and social systems and a has background in artificial intelligence, multi-agents systems and computational social science. He holds a M.S. and Ph.D. in Computer Science from the University of Illinois at Urbana-Champaign. Kiran's research has been marked by extensive interdisciplinary efforts that bring together the social and computational sciences, including developing models that explore the link between social structure (social networks, roles/hierarchy) with cognitive structure (how concepts are interrelated, cognitive consistency, confirmation bias) with respect to problems of information dissemination and attitude change.

Laura Swiler PhD has twenty-two years of experience in reliability and risk assessment. She started her career at Sandia working in reliability assessment, for weapons applications and nuclear waste repository assessment. For the past fifteen years, Laura has focused on sensitivity analysis, model validation, and uncertainty quantification. Some of Laura's research interests include: sensitivity analysis for high-dimensional inputs, model validation incorporating both experimental and simulation uncertainty, use of surrogate or meta-models for optimization and extrapolation, and calibration of model parameters in the presence of both experimental and model uncertainty.

Stephen Verzi PhD has more than 20 years of experience in software development, the last 15 years of which have involved in computational modeling of physical and biological systems. He has experience in System Dynamics, Agent-Based, Discrete-Event simulations and the underlying mathematics of dynamical systems. He is a member of the Complex Adaptive Systems of Systems (CASoS) Engineering group at Sandia as well as an adjunct professor at the University of New Mexico, teaching subjects such as computability and complexity theory and introduction to information theory. His current research interests include modeling human behavior and the population-wide consequences of such behaviors.

Vicente Romero PhD has been with Sandia National Laboratories for 30 years. He is in the UQ, V&V, and Credibility Processes department in the Engineering Sciences Directorate. He has a modeling background in optical, thermal, fluid, and structural systems, specializing in complex coupled systems and applications where statistical or stochastic behavior is important. Dr. Romero also has extensive experience in developing and applying optimization and uncertainty quantification techniques for model calibration, validation, and risk assessment and reduction in nuclear weapon systems subjected to stressing thermal-mechanical-electrical-radiation environments.