SAND2020-3737C

EXPERIMENTAL WARGAMES TO ADDRESS THE COMPLEXITY–SCARCITY GAP

Kiran Lakkaraju
Jason Reinhardt
Joshua Letchford
Jonathan Whetzel

Sandia National Laboratories
Livermore, CA, USA
{jcreinh,klakkar,jletchf}@sandia.gov

and Bethany L. Goldblum
Andrew W. Reddie

University of California - Berkeley
Berkeley, CA, USA
{bethany,areddie}@berkeley.edu

ABSTRACT

National security decisions are driven by complex, interconnected contextual, individual, and strategic variables. Modeling and simulation tools are often used to identify relevant patterns, which can then be shaped through policy remedies. In the paper to follow, however, we argue that models of these scenarios may be prone to the complexity-scarcity gap, in which relevant scenarios are too complex to model from first principles and data from historical scenarios are too sparse—making it difficult to draw representative conclusions. The result are models that are either too simple or are unduly biased by the assumptions of the analyst. We outline a new method of quantitative inquiry—experimental wargaming—as a means to bridge the complexity-scarcity gap that offers human-generated, empirical data to inform a variety of model and simulation tasks (model building, calibration, testing, and validation). Below, we briefly describe SIGNAL—our first-of-a-kind experimental wargame designed to study strategic stability in conflict settings with nuclear weapons. We then highlight the potential utility of this data for modeling and simulation efforts in the future using this data.

Keywords: wargames, experiments, national security, experimental wargaming, decision making

1 INTRODUCTION

Modeling and Simulation (M&S) tools play an important role informing our understanding of strategic decision making. Military planners, policymakers, and decisionmakers often use M&S to derive insights related to various national and international security challenges including military acquisitions to tactical level force structure to strategic decision making (Page 2016). In our work, we focus on the use of modeling and simulation tools to address challenges to strategic stability. Since the beginning of the Cold War, understanding

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

conflict escalation in the shadow of nuclear weapons is of particular strategic importance and the subject of extensive study and discussion. While much of the existing literature relies on qualitative sources to probe strategic problems, quantitative approaches such as M&S and empirical data analysis have an important, complementary role—particularly given the need for strategies that integrate analyses of security challenges across domains (Gartzke 2014). In the process, both traditional and cutting-edge M&S tools of conflict scenarios can be put to use (Roberts 2020, McNamara, Trucano, and Gieseler 2011).

Historically, approaches to modeling and analyzing conflict use both theoretical and empirical techniques. Game theoretic approaches, for example, often make assumptions about adversary knowledge and preferences and then create deductive models for adversary behaviors. These approaches, however, quickly become difficult to extend and computationally intractable for most realistic conflict scenarios that involve large numbers of players and contextual variables. As a result, much of the conflict modeling in the existing scholarly literature related to strategic stability focuses on two-player, one-shot nuclear deterrence and escalation games, often without accounting for uncertainty. For a survey and examples see (Zagare and Kilgour 2000); for exceptions see (Quackenbush 2016). Naturally, one must first question the assumptions made about the adversary that form the basis for such a model. Further, what these types of models suggest with respect to player behavior in dyadic, existential, single domain competitions may not extend to *n*-adic, non-existential conflicts across multiple domains.

For example, current theory for multi-actor conflict games largely centers on the so-called "truels" model, a three-way duel where each player can take a shot at either opponent, one-at-a-time, parameterized by a performance probability for each player (Kilgour and Brams 1997). Work on simple repeated and non-repeated variants of truels has led to surprising and non-intuitive optimal play strategies and game outcomes (Graham 2017). Existing work on simulating and solving multi-actor conflicts often uses agent-based simulations of theoretical strategies (Amengual and Toral 2006) or approximates multi-actor conflicts as a series of two-player conflicts (Hauert and Doebeli 2004). While an important starting point, these approaches are often computational intractable (Nash Equilibrium is PPAD-complete even with two players(Chen and Deng 2006), while other common solution concepts such as Stackelberg Equilibrium become NP-hard when a third player or imperfect information is added to the model(Letchford and Conitzer 2010)) and insufficient to address the complex, *n*-adic, multi-domain conflict space that is emerging (Roberts 2020). To address these gaps, analysts have turned to the empirical record to build and validate existing theory.

The examination of empirical data from historical conflicts represents an alternative approach to creating models of complex conflict scenarios. These approaches pursue data analysis techniques to identify behavioral trends and patterns that can provide insight on current scenarios. Datasets such as the Militarized Interstate Disputes (MIDS) collection (Maoz, Johnson, Kaplan, Ogunkoya, and Shreve 2019, Palmer, D'Orazio, Kenwick, and McManus 2019), for example, contain historical data on conflict between states. However, empirical data regarding nuclear crises are sparse and generalizations especially challenging given the wide range of factors that can influence conflict and competition outcomes. Moreover, the changing nature of technology and geopolitics may mean that future conflicts and strategically competitive situations may be less likely to resemble available historical analogs (Goldblum, Reddie, and Reinhardt 2019).

Overall, the underlying factors influencing strategic stability are varied and involve complex interaction that makes theoretical modeling difficult. On the other hand, empirical data concerning international crises to inform empirical modeling remains sparse. We describe this phenomenon as the “complexity-scarcity gap."

To address this gap, We suggest that synthetic, large-*n*, human-generated data derived from experimental wargames offer two valuable contributions to existing modeling and simulation efforts. First, this data might be used to inform efforts to build new conflict models (Reddie, Goldblum, Lakkaraju, Reinhardt, Nacht, and Epifanovskaya 2018). Second, this data might be used to calibrate, validate, and test existing models. Our experimental wargaming platform, described in greater detail below, focuses on scenarios that are difficult to model from first principles and for which scholars have a lack of data. Below, we argue that the data generated from experimental wargames can be used to (1) inform models of historical, current, and future conflict scenarios; (2) "learn" conflict behavior models from the data set; and (3) calibrate and inform M&S approaches.

Using experimental wargames as a source of human-generated data represents a new contribution to the existing M&S literature. Experimental wargames, are applicable to a range of conflict and competition scenarios including grey-zone conflict, proliferation challenges, and economic or treaty negotiations in addition to escalated conventional and nuclear conflict. Experimental wargames are also applicable to cooperative scenarios such as alliance building, consensus making, and coordinated strategy development. Importantly, the experiments can be performed on a system of multiple actors rather than just simplified, two-actor interactions.

Synthetic data from experimental wargames can be useful in the process of creating, calibrating, and validating models and simulations. Agent-based models (ABM) (Macal and North 2010), for example, would benefit from experimental wargames by gaining data to help in validation.

(Rand and Rust 2011) outline four steps to ensure rigor in validation of an ABM: Micro-face validation to evaluate how the model compares to the real world; Macro-face validation where processes and patterns from the model are compared to the real world; Empirical input validation to evaluate the data used as inputs to the model and Empirical output validation to compare the output of the model to the real world. For conflict related scenarios, especially when involving emerging technologies, there are limited data that can be used to conduct empirical output validation. Experimental wargames can provide these type of data to allow for stronger validation and assessment of the models capabilities.

(Smith and Rand 2017) discuss how controlled behavioral experiments with human subjects can inform ABMs by addressing the micro-face and empirical input validation steps. Experimental wargames, as we discuss in Sec. 2, emphasize experimental principles and can similarly serve as a way of informing the micro-face and empirical input validation steps.

Machine learning techniques may also be used to render conflict and competition behavior models from a combination of empirical and synthetic data sources. As an example of a machine learning algorithm applied to conflict policy, (Reinhardt 2018) which applied algorithms for solving Interactive Partially Observable Markov Decision Processes (IPOMDPs) to nuclear deterrence and conflict escalation dynamics to better understand the impacts of stockpile capabilities on antagonist strategies. The policy solutions are represented using an optimized finite state controller that represents an adaptive and learning adversary, with a rich set of conditional behaviors. These methods have distinct advantages in that they are compact, interpretable, include adaptive adversary models, capture complex behavior patterns in a variety of formats, and account for conflict escalation and a changing conflict environment. Synthetic data from experimental wargames can be used to calibrate these types of models (as well as other types).

In the section below, we describe our first-of-a-kind experimental wargaming platform, SIGNAL, and highlight some of the identified patterns of behavior. We conclude with a discussion of our broader vision of generating informed conflict models that integrate insights from synthetic data, empirical data, and behavioral principles to generate theoretical insights.

2 EXPERIMENTAL WARGAMES

Wargames have a long and rich history of use in a military context to instruct and inform decision makers and as a tool for exploring novel scenarios or proximate crises (Perla and Curry 2011). A large number of wargaming designs and platforms exist varying along multiple dimensions including level of abstraction, level of human interaction, context (strategic vs. tactical), and purpose (Sabin 2014, Marston 2016).

Seminar-based wargames (Development and Centre 2017), for example, represent a common type of strategic wargame (College 2015) in which multiple teams represent real-world states in a competitive environment. The teams are given a scenario and asked to describe the actions that they would take. The scenario and responses are often open-ended and an adjudicator ("white cell") interprets the actions and their consequences to move the game from one round to the next. In this type of game, the white cell determines the impact of the actions and injects events that fit the purpose of the wargame. Data collection from seminar-based wargames can be difficult as actions are often in verbal or text form, require effort to code, and are subject to potential coder and inter-coder biases. Executing a seminar-based wargame with elite players (which is often deemed desirable) is expensive, in both time and resources. These limitations make running seminar-based wargames at scale difficult, limiting the utility of these games for collecting data and rigorously testing and evaluating policy and strategic options. Nonetheless, these games can afford valuable insights; see (Pauly 2017) for instance.

Experimental wargaming focuses, instead, on generating a large amount of data about a specific scenario from many playthroughs—-complementing both existing wargames and other analytical methods. Our experimental wargaming methodological approaches emphasizes the following:

Repeatable:	The same game, under the same conditions, can be played many times by different players.
Controllable:	Variables of interest should be clearly and carefully controlled.
Instrumented:	Quantitative data on specific actions and outcomes can be captured easily.
Neutral:	Experimenters remain uninvolved with the data gathering, reducing bias.
Sufficient Realism:	Games create an appropriate abstraction of scenarios of interest (i.e., "distillation games" (Perla, Markowitz, and Weuve 2005)) with a focus on research questions and experimental design considerations.

Experimental wargaming allow the use of a control-treatment paradigm to assess the effects of specific conditions of interest. For example, analysts may assess the effects of various military capabilities on an idealized crisis. In a "control-treatment" paradigm, a group of subjects is assigned to the control condition and another group is assigned to the treatment condition. The control and treatment conditions are exactly the same except for one variable of interest. The primary analysis consists of observing and understanding changes in behavior between the conditions, which are assumed to be caused by the difference in the conditions.

One concern with studying behavior in a gaming environment is whether players act in a similar manner within a game world as in the real world. More broadly, this question ties to concerns about the external validity of behavior in a game. Several surveys have shown that game participants are demographically diverse and exhibit varying motivations (Kahn, Shen, Lu, Ratan, Coary, Hou, Meng, Osborn, and Williams 2015, Yee, Ducheneaut, Nelson, and LIkarish 2011, Brown 2018), mitigating potential sample bias. Case studies including the massively multiplayer online game, Eve Online, highlight insider threat activities painstakingly planned and executed over many months (Groen 2016).

As is the case throughout social science research, the goal is not to use an experimental wargame to predict behavior, but rather to identify a change in behavior conditioned on a change in experimental conditions. The results from an experimental wargame, similar to any experiment, illustrate the connection between the inclusion of the treatment condition and the resulting behavior. For the purposes of an internal validity, it is not necessary for players to act in a similar manner as in the real world as we are comparing two sets of conditions within the same game.

In using experimental wargames to understand the real world, the question naturally arises as to whether non-experts will act in the same way as military officers, national security professionals, or senior academics. We view this as an empirical question that should be tested rather than a limitation of the method.(Reddie, Goldblum, Reinhardt, Lakkaraju, Epifanovskaya, and Nacht 2019) Existing studies show that both non-experts and experts exhibit cognitive biases, but these biases may be different (Sheffer, Loewen, Soroka, Walgrave, and Sheafer 2018, Mintz, Redd, and Vedlitz 2006). Even if the decision making is significantly different for these two groups, the data can still be useful to highlight the space of strategies and behaviors that can be expressed.

3 SIGNAL: AN EXPERIMENTAL WARGAME TO STUDY STRATEGIC STABILITY

Now, we turn to a specific iteration of an experimental wargame, SIGNAL. Below, we outline the potential use of SIGNAL data to inform various M&S methods.

SIGNAL was designed to study conflict escalation in the context of varying nuclear weapon capabilities. It was developed as both a board game and an online game. The map used in the board game scenario is shown in Figure 1; the online game map is similar. It includes three countries, Orange, Purple, and Green, two of which (Purple and Green) are nuclear armed, while the third (Orange) has only conventional weapons. SIGNAL was designed without real-world countries to minimize players emphasizing pre-existing cultural biases. Players in SIGNAL have military and economic capabilities. Play takes place over a series of rounds where players may signal threats, execute actions, and negotiate with other players. Given the theorized importance of signaling in nuclear deterrence relationships, this first phase represents an important methodological contribution to wargame design. Players are assigned explicit goals to maintain territory, grow population, and grow resources and are evaluated on how well they have performed at each of these categories relative to the other players. SIGNAL was designed so that players can win through many means, including without conflict. Each round starts with an asynchronous phase where players can use trade and negotiation to shape the play of the other players. In this phase, players can commit to potential actions and locations via public staging to enable the option to execute these in the following phase. The second phase has a structured turn order where players may execute their staged actions at the previously staged locations. This pairing of phases requires that players reveal information about their intentions to other players prior to any actions taking place (in a way that allows for bluffing and misdirection) and enables players to use the capabilities that are available to them in that instance of the game both directly (to impact the positions of the other players) and indirectly (as threats for deterrence and coercion). Players are allowed to communicate with each other during the game, through face-to-face discussions in the board game and text chat in the online game.

Data were gathered on player signaling, actions, trade behavior, chat data (in the online game), and demographic characteristics. For more details about the game, see (Goldblum, Reddie, and Reinhardt 2019). The online version of SIGNAL can be found at signalvideogame.com.

Our goal with SIGNAL is to gather data that can be used to inform M&S. Thus, it was important to design SIGNAL with a variety of strategies, but we are not fundamentally interested in the game-theoretic properties of SIGNAL. SIGNAL provides data about a national security relevant scenario that we have minimal real world data on.

3.1 Applications of Experimental Wargaming Data to Modeling and Simulation

From SIGNAL, we collect extensive data about the process and progress of conflict, from initiation to resolution. These data allow for better calibration of conflict escalation and (deescalation) models. As

Figure 1: SIGNAL game board (board game version)

an example, we identify "conflict classes" within the data and show how these can be used to explore the evolution of conflict intensity nuclear-relevant scenarios.

The analysis shown here was based on the data collected from four board game events, comprising 44 games between December 3rd, 2018 to April 3rd, 2019. All data were recorded by the participants themselves as they were playing the game. These worksheets were collected from the participants and encoded into the data set.[1]

The data set can be parsed into discrete, regular time periods and the set of events occurring in that period can be categorized into classes that describe the nature of the game at that point, called conflict classes. For the purposes of this analysis, conflict classes are defined on a round-by-round basis. Specifically, conflict classes categorize the nature of the conflict based on a logic that depends on the combination of actions taken by all players within a round. This provides a convenient set of sequential categorical data that can be used for analysis.

The pattern of conflict occurring within the game is described through a sequence of conflict classes. Figure 2 illustrates an example conflict class sequence. The game begins in PEACE, indicating that no players have taken any actions that are explicitly threatening towards other players. Infrastructure construction, alliance building, economic trades, or verbal discussions may have occurred, but no military actions were threatened or taken. In the second round, at least one player has made a conventional threat (CONV THREAT), meaning that they have prepared to act militarily, and perhaps even signaled a willingness to take that action. However, no nuclear weapons were invoked or used, and no conventional actions were actually taken. In the third round, at least two players took conventional actions, categorized as a conventional exchange (CONV EXCHANGE), but no nuclear actions were taken. In the fourth round only a single player has taken a conventional action, categorized as a conventional action (CONV ACTION). In the fifth and final round of this example, multiple players employed nuclear weapons leading to a conflict class denoted as a nuclear exchange (NUC EXCHANGE).

Conflict classes can be defined based on the analytic question of interest (e.g., what was the path to nuclear use?). Analysis of conflict class data can help to identify trends in multipolar behavior through visualization and basic statistical techniques. While flexible, conflict class schema are simple enough to be easily analyzed

[1]At the time of publication we have collected approximatly 500 games using the online version of SIGNAL. This paper focuses solely on the board game data; analysis is currently being conducted on the online data.

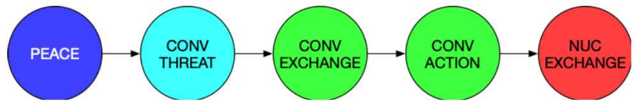

Figure 2: Conflict Class example.

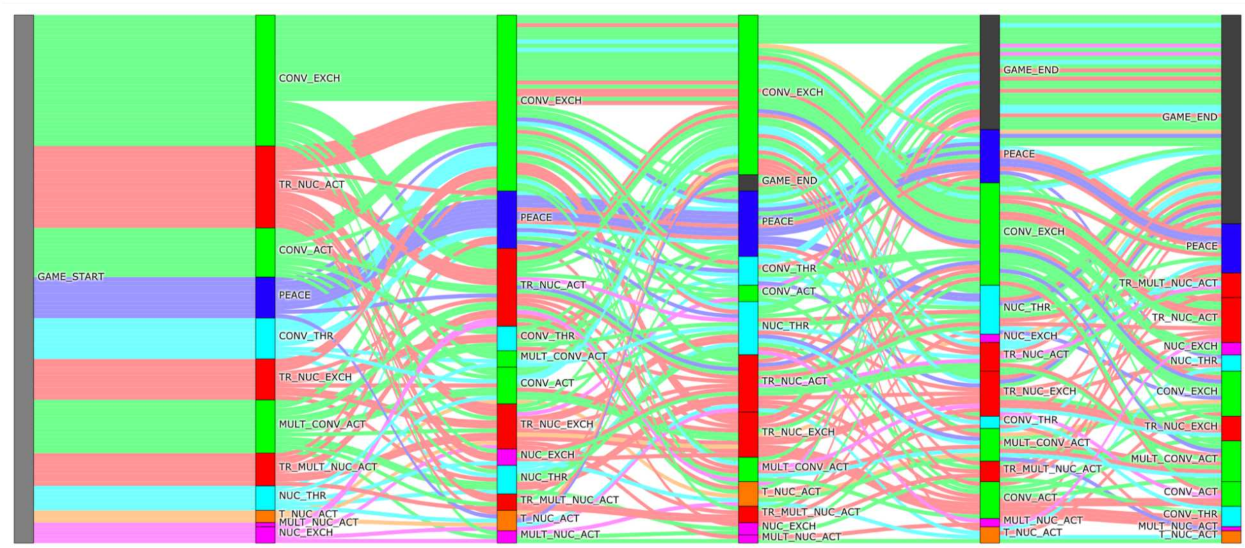

Figure 3: Once coded to a conflict class schema, the data set can be visualized as trajectories through the conflict class space, round by round.

and assessed, but can also obscure fine details in the game events if not carefully defined.[2].The coding of multiple actions into a single conflict class necessitate that some hierarchical order be enforced, so that some 'lesser' actions are obfuscated by the action that defines the class. For example, if a conflict class "NUC ACT" is defined such that any nuclear use in the round categorizes the round as NUC ACT, then any conventional action would be concealed. This could obscure the relevance of conventional actions in the assessment. Therefore, conflict classes should be thoughtfully defined and with careful interpretation of the results. The hierarchical conflict class structure used in this analysis is shown in Table 1.

One benefit of experimental wargaming lies in generating data about player behavior. Figure 3 shows a method of visualizing the conflict class data. From left to the right, we see successive conflict classes as we proceed through a game, where every column represents a new round. A line between two columns indicates a set of games in which there was a transition between the two conflict classes between two rounds.

Future M&S efforts can be calibrated by studying patterns of behavior within synthetic data. For instance, Figure 3 suggests that conflict early in the game can keep the game in a conflict state. This pattern would influence modeling choices (e.g., suggesting that agents with memory will be necessary, and furthermore, that a sufficient length of memory is included). This pattern and the patterns found via aggregating data into conflict classes could be used to inform empirical output validation or parameter calibration when building ABMs (Smith and Rand 2017).

[2]Data products from experimental wargames also address various units of analysis: player-level, dyad-level, temporal, or game-level

Table 1: Conflict Classes Used in This Analysis. Classes further down the table obscure classes above.

Peace	No players have staged any military cards for the resolution phase. Conventional Threat One or more players have staged conventional military action cards, and placed indicator tokens on opponent territory.
Nuclear Threat	One or more players have staged nuclear action cards, and placed indicator tokens on opponent territory.
Conventional Action	One player has played a single conventional military action card.
Multiple Conventional Action	One player has played multiple conventional action cards.
Conventional Exchange	More than one player has played conventional action cards.
Tailored Effect Nuclear Action	One player has played a single tailored effect nuclear action card.
Multiple Tailored Effect Nuclear Action	One player has played multiple tailored effect nuclear action cards.
Tailored Effect Nuclear Exchange	More than one player has played tailored effect nuclear action cards.
Traditional Effect Nuclear Action	One player has played a single traditional nuclear action card.
Multiple Traditional Effect Nuclear Action	One player has played multiple traditional nuclear action cards.
Traditional Effect Nuclear Exchange	More than one player has played traditional nuclear action cards.
Multiple Nuclear Action (Mixed)	One player has played a mix of traditional and tailored effect nuclear action cards.
Nuclear Exchange (Mixed)	More than one player has played a mix of traditional and tailored effect nuclear action cards.

4 DISCUSSION & FUTURE WORK

Historically, some strategic analyses informed by game theoretic models and game analytic methods have represented conflict in a relatively simple manner due to the complexity-scarcity gap — developing theoretical models that sufficiently represent the conflicts is difficult, and only limited data is available. In this paper we have argued that experimental wargames can bridge this gap by providing a means to gather large-*n* data from human subjects. These data can be used to calibrate, test, and validate models and simulations.

More broadly, we believe that the underlying components exist to build informed conflict models—models integrating synthetic and real data that incorporate behavioral science principles and attempt to correct for the impacts and biases of the experimental design. While traditional game analytic models are valuable tools for informing policy makers in strategic situations, they often rely on strong assumptions about actor behavior, game theoretic solution concepts, and only consider single models for adversaries. Informed conflict models can augment game analytic efforts by using observed behaviors from game play, corrected with behavioral science concepts, and experimental considerations to remove biases. Further, utilizing a plurality of potential adversary models to consider can allow analysts to develop strategies that are robust to a range of potential adversary types and actions. Thus, the informed conflict models are used as an augmentation to the classic game theoretic actor models, rather than a replacement. Figure 4 illustrates our philosophy concerning how conflict models informed by synthetic data might contribute to policy analysis.

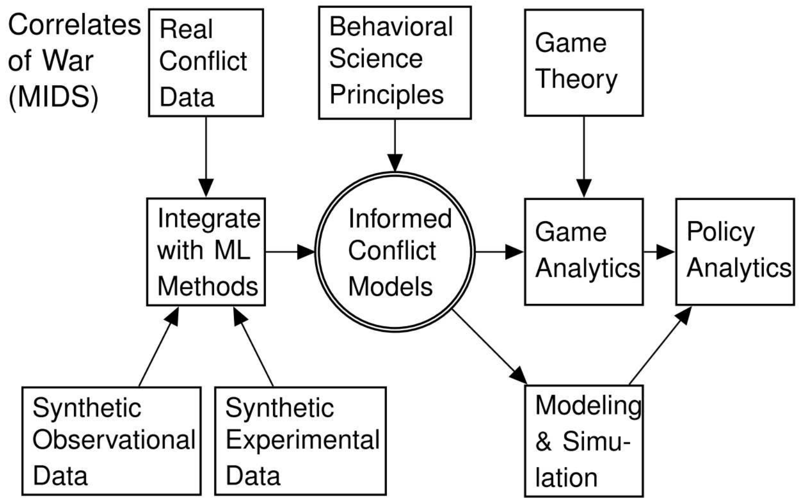

Figure 4: A vision for developing informed conflict models.

In this paper we have focused on using experimental wargames to generate synthetic experimental data. We can also leverage existing synthetic data sets, such as from Massively Multiplayer Online Games. These data sets are based on observation of games and often do not provide experimental controls, but are large and contain many instances of relevant conflict-related behavior (Lakkaraju, Epifanovskaya, Stites, Letchford, Reinhardt, and Whetzel 2019, Bainbridge 2007, Lakkaraju, Sukthankar, and Wigand 2018).

Advances in machine learning techniques provide a framework for integrating patterns from real conflict data with those found in synthetic data. In traditional machine learning approaches, an algorithm is applied to a data set from each problem domain, independently. This method does not leverage the potential similarities between problem domains that could enhance or speed up the learning process. Transfer learning techniques extract knowledge from one problem domain or setting and use that to enhance the learning process in a related domain or setting. In our case, transfer learning could be used to effectively leverage the extensive set of synthetic data to create models that can be applied to a small empirical data set, while still addressing potential differences in the underlying processes (Pan, Yang, et al. 2010). To better emphasize the patterns from real conflict data, transfer learning techniques can weight the different datasets differently.

The resulting machine learning algorithms yield informed conflict models that specify adversary behaviors in the examined scenario. These models can be built in a framework that includes a consideration of behavioral science principles and attempts to correct for the impacts and biases of the experimental design. Behavioral science encompasses the study of human behavior individually and within groups through scientific methods. Principles from behavioral sciences, such as prospect theory (Kahneman and Tversky 1990), can help inform our understanding and potentially explain patterns of behavior found in the data. Integrating these principles into machine learning techniques can enhance learning to identify relevant patterns of behvior that adhere to principles identified through extensive investigation in other fields. Techniques in machine learning have started to incorporate domain knowledge to enhance learning (Childs and Washburn 2019). While the focus has been on machine learning in physical systems, we believe there is the potential for these techniques to be used to integrate behavioral science principles as well.

Given the complex, interdependent strategic concerns facing decision-makers, M&S tools clearly offer a relevant method for analytical inquiry, but only if they can be appropriately tied to the real world. Experimental wargames represent a complementary tool to generate data that bridges the complexity-scarcity gap and that can provide data for both building future models and validating those that already exist.

REFERENCES

Amengual, P., and R. Toral. 2006, September. "Truels, or Survival of the Weakest". *Computing in Science Engineering* vol. 8 (5), pp. 88–95.

Bainbridge, W. S. 2007, July. "The Scientific Research Potential of Virtual Worlds". *Science* vol. 317 (5837), pp. 472–476.

Brown, A. 2018, May. "Younger men play video games, but so do a diverse group of other Americans". Fact tank, Pew Research Center.

Chen, X., and X. Deng. 2006. "Settling the complexity of two-player Nash equilibrium". In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pp. 261–272. IEEE.

Childs, C. M., and N. R. Washburn. 2019, September. "Embedding domain knowledge for machine learning of complex material systems". *MRS Communications* vol. 9 (3), pp. 806–820. Publisher: Cambridge University Press.

College, U. S. A. W. 2015. "Strategic Wargaming Series Handbook". Technical report, The United States Army War College.

Development, C., and D. Centre. 2017. "Wargaming Handbook". Technical report, Ministry of Defence - The Development, Concepts and Doctrine Centre.

Gartzke, E. 2014. "An Apology for Numbers in the Study of National Security…if an Apology is Really Necessary.". Technical Report No. 2, H-Diplo Forum.

Goldblum, B. L., A. W. Reddie, and J. C. Reinhardt. 2019. "Wargames as experiments: The Project on Nuclear Gaming's SIGNAL framework". *Bulletin of the Atomic Scientists*.

Graham, R. 2017. *The Mathematics of Various Entertaining Subjects: Research in Games, Graphs, Counting, and Complexity, Volume 2*. Princeton University Press.

Groen, A. 2016, April. "How EVE Online's greatest military leader once fooled the entire galaxy.". *PC Gamer*.

Hauert, C., and M. Doebeli. 2004, April. "Spatial structure often inhibits the evolution of cooperation in the snowdrift game". *Nature* vol. 428 (6983), pp. 643–646.

Kahn, A. S., C. Shen, L. Lu, R. A. Ratan, S. Coary, J. Hou, J. Meng, J. Osborn, and D. Williams. 2015, August. "The Trojan Player Typology: A cross-genre, cross-cultural, behaviorally validated scale of video game play motivations". *Computers in Human Behavior* vol. 49, pp. 354–361.

Kahneman, D., and A. Tversky. 1990. *Prospect theory: An analysis of decision under risk*. Rationality in action: Contemporary approaches. New York, NY, US, Cambridge University Press. Pages: 170.

Kilgour, D. M., and S. J. Brams. 1997. "The Truel". *Mathematics Magazine* vol. 70 (5), pp. 315–326.

Lakkaraju, K., L. Epifanovskaya, M. Stites, J. Letchford, J. Reinhardt, and J. Whetzel. 2019. *Online Games for Studying Human Behavior*, Chapter 17, pp. pp. 387–406. John Wiley and Sons, Ltd.

Lakkaraju, K., G. Sukthankar, and R. Wigand. (Eds.) 2018. *Social Interaction in Virtual Worlds*. Cambridge University Press.

Letchford, J., and V. Conitzer. 2010. "Computing optimal strategies to commit to in extensive-form games". In *Proceedings of the 11th ACM conference on Electronic commerce*, pp. 83–92.

Macal, C. M., and M. J. North. 2010, September. "Tutorial on agent-based modelling and simulation". vol. 4 (3), pp. 151–162.

Maoz, Z., P. L. Johnson, J. Kaplan, F. Ogunkoya, and A. P. Shreve. 2019, March. "The Dyadic Militarized Interstate Disputes (MIDs) Dataset Version 3.0: Logic, Characteristics, and Comparisons to Alternative Datasets". *Journal of Conflict Resolution* vol. 63 (3), pp. 811–835.

Colin Marston 2016. "Wargaming in the UK". Presented at the 2016 Connections UK Conference.

McNamara, L. A., T. G. Trucano, and C. Gieseler. 2011. "Challenges in Computational Social Modeling and Simulation for National Security Decision Making". Technical Report OSRD 2011 002, Defense Threat Reduction Agency.

Mintz, A., S. B. Redd, and A. Vedlitz. 2006, October. "Can We Generalize from Student Experiments to the Real World in Political Science, Military Affairs, and International Relations?". *Journal of Conflict Resolution* vol. 50 (5), pp. 757–776.

Page, E. H. 2016, August. "Modeling and Simulation, Experimentation, and Wargaming - Assessing a Common Landscape". Technical Report 16-2757, The MITRE Corporation.

Palmer, G., V. D'Orazio, M. R. Kenwick, and R. W. McManus. 2019, 07. "Updating the Militarized Interstate Dispute Data: A Response to Gibler, Miller, and Little". *International Studies Quarterly*. sqz045.

Pan, S. J., Q. Yang et al. 2010. "A survey on transfer learning". *IEEE Transactions on knowledge and data engineering* vol. 22 (10), pp. 1345–1359.

Pauly, R. 2017, February. "Nuclear Weapons in Wargames: Testing Traditions and Taboos". SSRN Scholarly Paper ID 2917628, Social Science Research Network, Rochester, NY.

Perla, P., and J. Curry. 2011. *Peter Perla's the Art of Wargaming a Guide for Professionals and Hobbyists*. Lulu.com.

Perla, P. P., M. Markowitz, and C. Weuve. 2005. *Game-Based Experimentation for Research in Command and Control and Shared Situational Awareness*. CNA.

Quackenbush, S. L. 2016, July. "Not Only Whether but Whom: Three-Party Extended Deterrence". *Journal of Conflict Resolution*. Publisher: Sage PublicationsSage CA: Thousand Oaks, CA.

Rand, W., and R. T. Rust. 2011. "Agent-based modeling in marketing: Guidelines for rigor". *International Journal of Research in Marketing* vol. 28 (3), pp. 181 – 193.

Reddie, A. W., B. L. Goldblum, K. Lakkaraju, J. Reinhardt, M. Nacht, and L. Epifanovskaya. 2018. "Next-generation wargames". *Science* vol. 362 (6421), pp. 1362–1364.

Reddie, A. W., B. L. Goldblum, J. Reinhardt, K. Lakkaraju, L. Epifanovskaya, and M. Nacht. 2019, March. "Applying wargames to real-world policies-Response". *Science* vol. 363 (6434), pp. 1406–1407.

Reinhardt, J. C. 2018. *A probabilistic analysis of the risk of nuclear deterrence failure*. Ph. D. thesis, Stanford University.

Roberts, B. 2020. "It's time to jettison Nuclear Posture Reviews". *Bulletin of the Atomic Scientists* vol. 76 (1), pp. 31–36.

Sabin, P. 2014, June. *Simulating War: Studying Conflict through Simulation Games*. Reprint edition ed. London Oxford New York New Delhi Sydney, Bloomsbury Academic.

Sheffer, L., P. J. Loewen, S. Soroka, S. Walgrave, and T. Sheafer. 2018, May. "Nonrepresentative Representatives: An Experimental Study of the Decision Making of Elected Politicians". *American Political Science Review* vol. 112 (2), pp. 302–321.

Smith, E. B., and W. Rand. 2017, December. "Simulating Macro-Level Effects from Micro-Level Observations". *Management Science* vol. 64 (11), pp. 5405–5421.

Yee, N., N. Ducheneaut, L. Nelson, and P. LIkarish. 2011. "Introverted Elves & Conscientious Gnomes: The Expression of Personality in World of Warcraft". In *Proceedings of CHI 2011*.

Zagare, F. C., and D. M. Kilgour. 2000, October. *Perfect Deterrence*. Cambridge, UK : New York, NY, Cambridge University Press.

AUTHOR BIOGRAPHIES

KIRAN LAKKARAJU is a Senior Member of the Technical Staff at Sandia National Laboratories, California. He holds a PhD in Computer Science from the University of Illinois at Urbana-Champaign. His research interests lie in wargames and agent-based modeling. His email address is klakkar@sandia.gov.

JASON REINHARDT is a Distinguished Member of the Technical Staff at Sandia National Laboratories. He received a PhD in Decision and Risk Analysis from Stanford University with a focus on applications to national security and deterrence. His research interests are in probabilistic analysis, game theoretic models, and experimental wargaming. His email address is jcreinh@sandia.gov.

JOSHUA LETCHFORD is a Senior Member of the Technical Staff at Sandia National Laboratories, California. He holds a PhD in Computer Science from Duke University. His research interests include game theory, optimization and wargaming. He is reachable at jletchf@sandia.gov.

JON WHETZEL is a Principal Member of Technical Staff at Sandia National Laboratories. He holds an M.S. in Computer Science from Texas A&M University, His research interests include artificial intelligence and machine learning, with applications to serious games, complexity theory, and cognitive science. His email address is jhwhetz@sandia.gov.

ANDREW REDDIE is a postdoctoral fellow at the University of California, Berkeley. He currently serves as deputy director for the Nuclear Policy Working Group and as a researcher for the Department of Nuclear Engineering, Goldman School of Public Policy, Center for Long-Term Cybersecurity, Berkeley Asia-Pacific Study Center, and the Project for Nuclear Gaming. He can be reached at: areddie@berkeley.edu.

BETHANY GOLDBLUM is an Associate Research Engineer in the Department of Nuclear Engineering at the University of California, Berkeley. She is a strategic thinker, data sense-maker, and team leader focused on nuclear security science and policy. Her email address is bethany@nuc.berkeley.edu.

5 ACKNOWLEDGEMENTS

This work was supported by a grant from the Carnegie Corporation of New York through their International Peace and Security Program. The authors recognize support from the Nuclear Science and Security Consortium at the University of California, Berkeley, under contract DE-NA0003180; the Center for Global Security Research at Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344; and Sandia National Laboratories, 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. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States government; Lawrence Livermore National Security, LLC; or Sandia National Laboratories. The authors would also like to acknowledge members of the Project on Nuclear Gaming past and present.