Received: 20 July 2020 | Revised: 5 January 2021 | Accepted: 5 January 2021

DOI: 10.1002/ffo2.64

ORIGINAL ARTICLE

WILEY

The simulation manifesto: The limits of brute-force empiricism in geopolitical forecasting

Ian S. Lustick | **Philip E. Tetlock**

University of Pennsylvania, Philadelphia, PA, USA

Correspondence
Ian S. Lustick, University of Pennsylvania, Philadelphia, PA, USA.
Email: ilustick@sas.upenn.edu

Funding information
This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. The authors thank Miguel Garces and Louise Lu for their invaluable contributions to this project.

Abstract

Intelligence analysis has traditionally relied on inside-view, case-specific modes of thinking: why did this actor—say, the USSR—do that and what might it do next? After 9/11, however, analysts faced a vastly wider range of threats that necessitated outside-view, statistical modes of reasoning: how likely are threats to emerge from actors of diverse types operating in situations of diverse types? Area-study specialists (who staffed most geopolitical desks) were ill-equipped for answering these questions. Thanks to advances in long-range sensing, digitization, and computing, the intelligence community was flooded with data, but lacked clear ideas about how to render it relevant. Empiricism, whether grounded in deep inside-view knowledge of particular places or broad outside-view knowledge of statistical patterns across the globe, could not and cannot solve the problem of anticipating high-impact, rare events, like sneak attacks and pandemics. Contingency planning for these threats requires well-calibrated conditional forecasts of the impact of policy interventions that in turn require synthesizing inside- and outside-view analytics. Such syntheses will be best achieved by refining computer simulations that permit replays of history based on the interplay among initial conditions, chance, and social-science models of causation. We offer suggestions for accelerating the development and application of theory-guided simulation techniques.

KEYWORDS

big data, computer simulation, geopolitical forecasting, inside-view, intelligence analysis, outside-view, scenario-building

1 | OVERVIEW: TWO SHOCKS AND THEIR ANALYTIC CONSEQUENCES

President Franklin Roosevelt famously declared that December 7, 1941, the date of Japan's attack on Pearl Harbor, would "live in infamy." He meant that the surprise onslaught against the American fleet was so dishonorable that Japan would forever suffer from association with it. But what shook the American public to its core was not the perfidy of Imperial Japan but the shock of so severe a challenge to the country's ability to defend itself and the lack of foresight it seemed to reveal. For decades, controversy would rage over whether the defeat at Pearl Harbor could have been avoided. Although some charged Roosevelt with knowingly allowing the attack to help his campaign to enter the war against Germany, most critics treated it as a failure of intelligence analysts who had accurate reports of the impending attack but did not turn them into the timely warnings that the political echelon would heed (Wohlstetter, 1962).

In the wake of the disaster, Roosevelt shut down the Office of the Director of Information, which had only been in existence since July 1941. Having apparently failed in its mission to coordinate disparate channels of information, it was replaced in June 1942 by the Office of Strategic Services (OSS), which after WWII

developed into the Central Intelligence Agency. In the midst of a plainly existential conflict with the Axis powers, the mission of the OSS was clear—to penetrate German, Italian, and Japanese political-military apparatus, to discover motives, capabilities, and intentions, and to engage in clandestine counter-intelligence and sabotage operations.

On September 11, 2001, the United States was the target of another surprise attack, al-Qaeda's destruction of the twin towers in New York and a portion of the Pentagon, along with the downing of a passenger plane in Pennsylvania. Washington's reaction was again to go to war, but identifying the enemy was much more difficult. Al-Qaeda itself was attacked in Afghanistan and elsewhere but the "Global War on Terror" was, as the name suggests, a broad-spectrum mobilization against a kind of tactic, and the threat of its use by what came standardly to be termed "the universal adversary" (Lustick, 2006, p. 124). The country was believed to be facing an "all azimuth threat" (Lustick, 2006, pp. 4, 96). Attacks could arise from anywhere, anytime, against any target. Both law enforcement and intelligence agencies were widely blamed for the failures of coordination, information sharing, imagination, and asset deployment that, it was believed, not only enabled the 9/11 attack but raised the risk of other unprecedented types of attacks.

To increase its capabilities to prevent or mitigate such attacks, the government moved quickly, once again, to reorganize. It replaced the Central Intelligence Agency with the Office of the Director of National Intelligence as the organization with overall responsibility for espionage, analysis, and clandestine operations. While still *primus inter pares*, the CIA now reported to ODNI, as did an alphabet soup of other agencies, including NSA, DIA, INR, and the intelligence units of the armed services and of the Office of the Chief of Staff.

More important than these organizational changes, the end of the Cold War and the post-9/11 launch of the Global War on Terror dramatically changed the overall mission assigned to the intelligence community (IC) and the scale of the resources allocated to it. Intelligence agencies played major roles in U.S. wars in Afghanistan and Iraq. But those were only particular fronts in a much wider and amorphous effort to gather information about every conflict and potential conflict in the world that could breed the sort of asymmetric but devastating attack visited on the country on 9/11.

The changes in the IC's responsibilities, resources, targets, expectations, and organizational structure have been widely discussed by both academics (Betts, 2007; Jervis, 2010; Lustick, 2006; Tetlock & Mellers, 2011) and insiders (Hayden, 2017, 2019; Hitz & Weiss, 2004; Snowden, 2019; Steinberger, 2020). Both experts and the general public are also aware of revolutionary changes in the role played to achieve these goals via surveillance, analysis, and operations, by computing, cyber techniques, satellite technology, drone use, and artificial intelligence. Largely hidden from view, however, have been the inadequacies of both theory and existing analytic techniques for answering the new kinds of questions the IC must address.

Drawing on traditional anthropological and contemporary cognitive psychology categories to distinguish two approaches to geopolitical forecasting, we will offer a matrix classifying the challenges intelligence analysts face according to the availability of data and of theory bearing on the target problem. Using this matrix, we will locate strategic issues confronting the IC and document the limitations of conventional scenario-building tools for addressing them. We then demonstrate the power of theory-guided computer simulation to overcome these limitations—and to leverage social science knowledge for systematic mapping of possibility and plausibility spaces and for testing counterfactual and conditional claims about the effects of alternative policies.

During World War II and the Cold War the OSS and the CIA focused mainly on what Kahneman (2011) calls "the inside view" or anthropologists call the "emic" approach. Both are "particularizing" modes of thought that involve identifying the internal elements of a system and their interactions, with little interest in comparison classes of similar systems. Using clandestine techniques, agents and analysts would gain privileged access to the motivations, capabilities, and intentions of precisely specified targets. Drawing heavily on highly classified information, analysts would sharpen accounts of what the enemy could do, wanted to do, and intended to do. The standard questions they sought to answer took the following form.

1. How likely was Germany to launch an attack of a particular type on a particular front with particular weapons and forces?
2. How long could Russian forces hold out against German pressure against Stalingrad?
3. What were the chances of Rommel's Afrika Corps taking Egypt from the British and threatening the core of the Middle East?
4. What tactics would the Japanese use to resist American island hopping?
5. How would the Soviet Union react in Cuba to American strategic nuclear superiority?
6. What kind of support would the Soviet Union provide to the North Vietnamese in the face of sustained American bombing?
7. During the "War of Attrition" between Israel and its Arab neighbors in 1969 and 1970, how likely was it that the Soviet Union would supply Egypt with enough anti-aircraft missiles and Soviet pilots to defeat Israel in the air over Suez?
8. If the Israelis attacked the Egyptian Third Army at the end of the 1973 War, would that trigger Soviet armed intervention?
9. With what likelihood and over what time frame would economic underperformance in the Soviet Union lead to the demise of the Communist bloc?

The products of intelligence analysis, as consumed by decision makers, were thus answers to questions about particular actors in particular places and times. Answers were expected in the form of informal or formal probability judgments about how events would unfold and why, along with analysis of what steps could be taken, when, where, and how, to further American objectives or thwart threats. Those deemed most capable of carrying out these tasks were generalists and area-study specialists—men and women with exposure to the "great game" of international affairs, with military experience and

TABLE 1 "What-World-Am-I-In?" Matrix

	Plentiful Data	*Scarce Data*
High confidence in ontological priors about entities and causation	**1 Close-to-optimal blend of data and theory:** Possible to integrate outside/statistical views with inside/causal-theory views by testing hypotheses in environments that provide timely accuracy feedback. Potential for accurate forecasting limited only by the irreducible randomness of the world	**2 Dominance of speculation over data:** What-if scenarios based on theoretical or narrativist hunches about causation not grounded in data. Little potential for forecasting but some for contingency planning. Risk of dogmatism and high-assertion-to-evidence ratios
Low confidence in ontological priors	**3 Dominance of number-crunching over thinking.** Use big data and powerful statistical tools to identify patterns and generate hypotheses. Potential for forecasting but serious challenges in linking statistical evidence to policy questions. Risk of confusion and overload	**4 Unsustainable uncertainty**: Temptations to trust prophets or embrace intellectual fads that promise speedy reduction of ambiguity. Prioritize redefining the problem so that quadrant 2 or 3 solutions apply

expertise, or who had immersed themselves in the cultures and societies that were the targets of intelligence operations or the arenas within which competition with those targets was taking place.

After 9/11, however, with the vast expansion in the range of threats in the Global War on Terror, it was no longer possible for the IC to imagine succeeding in its mission by sticking to its traditional emphasis on the emic, or "inside-view." The idea of a "universal adversary" meant that fiendishly obscure but catastrophic threats could arise from anywhere anytime, and implied the need for orders of magnitude more data than could be generated by traditional clandestine penetration of targets and collecting and classifying case-specific information.

The response to this epic challenge was energetic and infused with a determination that the United States would never again be stunned by a Pearl Harbor or 9/11-style surprise. Flush with funds, the IC reached out to the scientific community. The collaboration was imagined as akin to the Manhattan Project during WWII that harnessed the best brains, wherever they were, and gave them deep-pocket budgets for the development and production of the first atomic bomb. Close involvement in intelligence matters of civilians with low level or no security clearance was facilitated by the gradual realization that most relevant data were available in open, rather than classified, sources.

Encouraged by progress in the social sciences, statistics, and by technological advances in remote sensing, supercomputing, and artificial intelligence/machine learning, the IC launched a host of new programs to find techniques for an outside-view or "etic" approach to solve "hard," "DARPA hard," (DARPA: Defense Advanced Research Projects Agency) or "wicked" problems—problems characterized by "deep uncertainty." Whether sponsored by the CIA, DARPA, ODNI (Office of the Director of National Intelligence), or IARPA (Intelligence Advanced Research Projects Activity), such programs were designed to develop collection-and-analysis techniques capable of surveying the globe, along a great number of dimensions, and produce forecasts of emergent threats no matter how unusual or even "unique" their profile.

But with an intricately interdependent world now defining the field within which threats of any form could emerge at any time, the priority task in the post-9/11 era became cognitive triage: coping with immense amounts of data combined with demands to think "outside-the-box" with sufficient imagination to have anticipated threats as diverse as 9/11 terrorism, cyber-hacking, bio-warfare, nuclear proliferation, and social-media manipulation. The imperative became gauging the plausibility, scale, and urgency of thousands of threat vectors. The government poured immense resources into intelligence and counterterrorism. In return, it expected not only reliable forecasts but "actionable intelligence"—guidelines for actions to pre-empt or mitigate threats.

This is a qualitatively harder problem than the IC had ever faced--a "wicked" problem, with no strong theory to guide where to look, no boundaries to limit the range of data that could be deemed relevant, and no opportunities to safely learn from trial and error. Such problems fall into quadrant 4 of the "what-world-am-I-in?" matrix—a taxonomy of epistemic opportunities and predicaments that is organized around the availability of credible inside and outside views. In Table 1 the rows of the matrix are defined by answers to "How much do analysts know about causal origins of the threat?" and the columns by answers to "How much do analysts know about the statistical patterns of types of threats across types of entities?"[1]

When we understand the specific threat and have plentiful comparative data, we reside in comfortable quadrant 1. Here are the familiar intelligence-analysis problems: known enemies whose profiles we have observed, openly or clandestinely, for a long time: a new weapon deployment by an adversary that changes the vector and scale of military threat; the interpretation of geophysical indicators of underground nuclear explosions revealing weapon-development strategies. For such problems, the IC already has most of what it needs to test policy-relevant hypotheses derived from models grounded in sound theory.

But when we lack good theories about the threat and good data to evaluate our guesses, we find ourselves in uncomfortable quadrant 4, where prophets flourish but analysts flounder. If analysts are going to have anything useful to tell policy makers, they must figure out, fast, an escape route—either by moving left, to quadrant 3, (find comparison-class data) or moving up to quadrant 2 (find sound theories). As we shall see, for technological, political,

and sociology-of-knowledge reasons, the IC has chosen to invest vastly more effort into moving from quadrant 4 to quadrant 3, where statisticians can construct outside views from huge datasets, than to quadrant 2, where we need social-science expertise to construct inside views from narrative or theory-driven schemas.

Quadrant 3 is the realm of the number-crunchers: sound theory is scarce but comparison-classes are plentiful. Here analysts can use standard statistical tools—OLS regression, computational linguistics, and machine-learning pattern recognition—to generate testable forecasts. But the risks of capitalizing on chance are enormous. For example, we might discover that bioterrorist threats are high when the price of lemons is low. Such correlations yield hunches about which relationships merit more attention, but we cannot mitigate bioterrorist threats by increasing the supply of lemons. Brute-force empiricism can inspire testable hypotheses about causality but not much beyond that.[2]

Quadrant 3 also raises another problem. When massive correlation matrices are plentiful, but theory scarce, analysts gravitate toward posing questions that data-analysis tools can answer. The result is techno-empiricism, a warping of inquiry reminiscent of the parable of the drunk who confined his search for his wallet around the nearest lamppost because that is where the light was brightest. In Quadrant 3, things are not quite that bad. The accuracy of some forecasts can be increased by extrapolating trends and matching current configurations to past patterns but the payoffs are circumscribed by the distance between the problem that analysts can solve and the problem that policy-makers want solved. Such techniques yield accurate probability forecasts for common events, like political stability in Singapore next year. But they cannot anticipate unlikely events—those of greatest interest—like political *instability* in Singapore a year from now. Nor, absent theories about the drivers of rare, high-impact events, can such analysis guide us in dealing with the challenges that the IC has been charged to prevent or mitigate.

Quadrant 2 has traditionally been the realm of gurus who do not need to fret over being ambushed by inconvenient facts.[3] High-fidelity data, that would let us winnow out theoretical wheat from rhetorical chaff—are just not available. So analysts must rely on their own judgment in picking from the many theories on offer in the marketplace of ideas. And, once picked, these theories morph into stylized facts. This blurring of empirical facts and theoretical opinion is less problematic in the mature sciences where we have strong theories. Simulations of electrical grid performance or supersonic fighters are almost as good as the real thing. But when the theories are weak, the blurring becomes profoundly problematic. Debates over how close we came to nuclear war during the Cold War arguably tell us far more about the debaters' assumptions about the robustness of nuclear deterrence than they do about the plausibility of specific close-call counterfactuals in the 1950s. 60s, 70s, and 80s (Tetlock, 2017, Chapter 5).

Our main purpose is to improve forecasting and decision making in quadrant 2. This will require combining substantive theory and computer simulation with an ambition and rigor that is commonplace in the physical sciences but has rarely been attempted for geopolitical problems.[4] The typical inferential engine used by national-security or foreign-policy analysts is not formal theory but hunches, explored imaginatively, about how the future might resemble episodes from the past, or exemplify favorite precepts or proverbs. Stories about the future produced in this way are scenarios, the default sensing technique in quadrant 2. The question we pose is whether theory-informed computer simulation enables the production of distributions of scenarios that are orders of magnitude more informative about the future than those that can be produced inside the heads of analysts. And the answer we propose is "yes."

In principle, producing scenarios with computer simulation or with human brains involves identical operations: an information processor constructs representations of initial conditions, causal covering laws and projected sequences of events condensable into coherent stories. Policy-makers, the consumers of analytic products, may not even know which stories are the product of analysts' intuitive efforts to imagine consequences or the product of a computer program running large numbers of trajectories. Either way, the resulting scenarios serve the vital function of surrogate data. Although we cannot hop into time machines that would let us collect actual observations of possible worlds and test our cause-effect claims in policy debates, we can at least approach the task of generating hypothetical observations in a transparent and rigorous fashion.

2 | SCENARIOS: BEYOND THOUGHTFUL IMPRESSIONISM

Sketching alternative stories of the future and using them to think through decisions in the present originated with Hermann Kahn's speculations about nuclear war, but gained credence and a standardized vocabulary with Pierre Wack's successful scenario work for Royal Dutch Shell in the 1960s and 1970s (Klein, 2003). Whether advanced to illustrate possibilities or to highlight boundary conditions, scenarios are presented as plausible narratives about what could happen (or would have happened, if attention is directed to counterfactual retrodiction) and why. Each is comprised of implicit forecasts based on hypothetical conditionals, such as exogenous changes in circumstances, strategies of adversaries, or policy decisions of those considering the scenarios. Attempts to formalize the technique have converged on ways of structuring conversations that elicit coherent alternative narratives from distinct "theoretical logics" or "intuitive logics," which is to say from commitments to alternative views of the drivers of change (Bradfield, Wright, Burt, Cairns, & Van Der Heijden 2005; Huss & Honton, 1987).

Practically speaking, scenarios tend to be colorful inside-view accounts of events as they could unfold if key causal drivers took on either lower or higher values. For example, experts in an analytic workshop might be asked to imagine possibilities conditional on lower or upper bound casualties for a pandemic—or lower or upper bounds on a leftward/rightward shift in governance. While policy makers, business planners, and intelligence analysts standardly use their own priors to guide these scenario thought

experiments, academicians often draw on formal theories or statistical models. For example, game theorists describe outcomes by imagining key players interacting according to payoff structures of games. Aggregate data analysts accomplish an analogous task by inferring patterns in the future similar to those measured in the past.[5]

The popularity of the scenario method is understandable. People like engaging narratives and endow them with credibility. And the richer the details, the easier it is to transport ourselves into the story (Tetlock, 2017). But scenarios can be no more reliable than the human minds generating them. And there are good grounds for suspecting that, though people often leave scenario exercises feeling they have learned a lot, there will often be serious mismatches between subjective reports of learning and objective metrics. The root problem lies in what cognitive psychologists call the illusion of explanatory depth (Keil, 2005). Unless called rigorously to account, people rather reflexively exaggerate their understanding of the workings of even simple, everyday systems, like bicycles and toilets, and the gap may well grow wider when people take on remote topics, like nuclear deterrence or monetary policy. Consider two of the many ways in which scenarios can overwhelm or befuddle us.

First, it is easy to imagine altering a causal antecedent: if there had been a Gorbachev-style leader in China in 1989 or a Deng Xiao-ping-style leader in Moscow in 1985. And it is then tempting to suppose that, all else equal, the Soviet Union would have put greater priority on economic liberalization and China would have put greater priority on political liberalization. But these thought experiments rest on an "all other things equal" assumption that makes them easy for humans to "compute," but also highly unrealistic because they ignore the systemic ramifications of altering an antecedent, a problem known as cotenability in the literature on counterfactual thought experiments (Tetlock & Belkin, 1996; Tetlock et al., 2006). For instance, imagining a Gorbachev-style of leadership emerging in China during the Tienanmen demonstrations of 1989 also requires imagining a very different type of Chinese Communist Party, which also requires… Similarly, imagining a Deng Xiao-ping-style leadership emerging in Russia in 1985 requires imaging a very different Communist Party of the Soviet Union, which also requires… The complexity of the ripple effects quickly becomes overwhelming—as does the temptation to ignore the complexity.

Second, the same enhancements of detail that make vivid scenarios compelling also make them, as a matter of logical necessity, less probable. Subsets of possibilities cannot be likelier than the sets from which they were derived. But Tversky and Koehler's (1994) support theory warns us to expect people to routinely commit this logical fallacy. The likelihood of a major flood in North America cannot be less likely than the likelihood of a flood triggered by terrorist sabotage of a dam in California—but the latter possibility comes much more readily to mind and it is also a much more promising premise for a movie script. Indeed, Tetlock (2017, Chapter 7) found that when one added up the probabilities that experts assigned to the logically exhaustive and exclusive possibilities implied by a scenario analysis, the sum often exceeded 1.0—and occasionally approached 2.0.

Scenario builders seldom attend to these problems of endogeneity, under-specificity, replicability, or to the sheer noisiness of an inference process in which people must make effect size estimates for interacting combinations of variables. It should not be surprising that consultants, whose paychecks hinge on customer satisfaction, prefer to exploit psycho-logics that inflate the subjective probabilities of unjustifiable but entertaining forecasts.

All of which leaves the analytic community in a conundrum. Forecasters need divergent thinking and imaginative scenarios but also the conceptual discipline to winnow out incoherent probability estimates of risks and opportunities. Some human beings are surely better than others at rising to this challenge of systematizing imagination and cobbling together alternative pictures of the future that blend what is known and what is possible. But we see it as a good cognitive-psychological bet that if researchers could automate the application of relevant theories to high-quality data, and model the state space of probable, plausible, and possible outcomes under randomly perturbed conditions, the system would out-perform the theorists themselves—and do so by virtue of its superior capacity to reduce noise in the information integration process (Kahneman et al., 2021). Theory-informed computer simulation is just such a technology, and although its availability and potential have not been widely appreciated by policy makers or strategic analysts, it has already been successfully applied to real geopolitical problems.

Thought experiments are essential for exploring the murky space of the possible, the plausible, and the probable in quadrant 2. Whether produced on wetware or software, in brains or on silicon chips, those who conduct them confront problems of quality control. But the difficulties human scenario-builders face, whether they draw on private rumination or structured interactions among multiple minds, are orders of magnitude more daunting than those facing analysts capable of charging computers to answer the same questions. Consider that any thought experiment requires inferences about unobserved events linked to unobserved chains of causal contingencies. Whether human analysts perform these thought experiments, or program a computer, the process is fundamentally the same, and so is the form if not the quality of the result: accounts of possible futures or possible pasts generated by imaginative extrapolation from beliefs about initial conditions interacting with the analyst's theories about how the world works. But the differences are enormous. When performed by human beings, the theories deployed are implicit or unknown and often an inconsistent jumble. The processes of translating those theories into specific claims are under-specified and the conclusions cannot be reliably replicated. When performed via theory-guided computer simulation, however, the theories must be explicit and consistent. Models translating those theories into specific claims are transparent and conclusions inferred from patterns in large numbers of scenarios generated are fully, if not necessarily easily, replicable.

It would be remarkable if two analysts working independently from the same complex information set converged on precisely the same scenarios—or that they could reproduce the exact thought processes underlying their results. Human judgment is just too

noisy—and our introspective access to our mental machinery too imperfect (Kahneman et al., 2021). Indeed, it is highly unlikely that even the same analyst, or group of analysts, if asked to devise scenarios responsive to specified conditions, would produce the same detailed accounts twice. Research has shown that a simple computer model, designed to capture the basic inference strategies of an expert in a domain, can quite reliably out-perform the expert from whom the model was entirely derived (Dawes, 1979; Kahneman et al., 2021). And there is little mystery about how the model emerges triumphant. It does not know anything beyond what the expert knows but it can apply the expert's rules of thumb with perfect reliability, something virtually no experts can do.

For all these reasons, human analysts should wonder whether the theories underlying their scenarios do not contain assumptions that, flushed into the open, would be contradictory. Exacerbating matters, these latent contradictions are often obscured by the narrative seductiveness of the scenarios—and the use of vague verbiage phrases, like "distinct possibility or "highly plausible" that readers translate into wide spans of probabilities.

In contrast, computer simulations are relatively transparent and can be perfectly replicable. Their formal algorithms are not open to human interpretation. The theories operationalized within their programs must be logically consistent. The requirement of cotenability is automatically fulfilled—and the incoherent probability judgments predicted by support theory avoided. By exposing trajectories arising from the same modeling platform to small random, exogenous perturbations, and using other Monte Carlo techniques, every scenario trajectory can be unique in its details, yet fully documented. Most crucial from a causation perspective, nothing can happen within any one computer-generated trajectory that is inconsistent with the rules governing all other trajectories.

Computer simulation thus holds the promise of vastly increasing the complexity, rigor, and number of disciplined thought experiments we can perform. When thousands of trajectories can be generated and systematically compared, simulations can map the state space of the future, identifying zones that are unlikely to be visited but contain high-impact outcomes. And given that each trajectory is governed by laws with traceable event trails and branching points, crucial precursors to threats and opportunities can be located. Herein lies the potential of theory-guided computer simulations for forecasting in quadrant 2 and for mitigating extreme impact events.

For it is not just outcomes but the trajectories traced by computer simulation that are fully replicable. Human authors of scenarios are incapable of specifying the staggering number of "transition rules" driving the outcomes they speculate about. And even within the domains of their distinctive expertise, the laws of social behavior at multiple levels of analysis are at best only partly understood. Whether produced by individual human brains or by interactions among humans, such stories about the future, and the likelihoods of events they imply, must include large doses of arbitrary, even idiosyncratic, decisions about which variables will dominate, in what combinations, in what sequence, and with what consequences. This again makes replication of human-generated scenarios virtually impossible.

We can practice simulation-enhanced thinking about the future by systematically using the language of counterfactuals. All claims about the future are analogous to historical counterfactuals: conditional statements about something that could or will happen under conditions not yet realized—and perhaps never to be realized. And given that most errors about the future cannot be identified definitively until the actual future becomes the past, a critical stance toward claims about the future cannot rest on categorical assertions of truth over falsehood. It requires thinking not about the likelihood of imagined trajectories within what is often tacitly assumed a quasi-normal distribution, but about differently shaped distributions of trajectories. In other words, what is needed is a map of the space of the future (or at least a map of a virtualized version of the future) showing how uncertainty is distributed—a map comprised of immense numbers of events, all consistent with the theories operationalized by the program that produces them, and each connected to others in causally specified patterns of interaction. These patterns depict accessibility to different zones within the space of traced futures, signaling degrees of likelihood for particular kinds of events, and offering opportunities to identify precursors and branching points.

3 | BRINGING CAUSATION BACK IN

Spaniol and Rowland (2019) have proposed that a consensus has emerged in the forecasting community: namely, scenario-building is best treated as an exercise in deploying and aggregating "intuitive logics." We agree but see this as a thin reed on which to base weighty decisions that could put whole nations and their populations at risk. Indeed, scenario planners have themselves long been aware of the lack of rigor of the most common approach for thinking about the future: BOGSAT (Bunch of Guys Sitting Around Talking). The problem has been partly addressed by scholars of group dynamics and management who have developed techniques aimed at improving the scenario-building process, including pooling independently elicited judgments, extrapolating trend lines, and looking at data through the lenses of alternative assumptions (Ramírez et al., 2013). But as Wright et al. (2013) point out, we cannot claim the intuitive logics approach has improved "strategic conversations," if by success we mean better decisions or outcomes.

Davis, Bankes, and Egner (2007) report an unusually systematic attempt to discipline scenario production, which they call "Massive Scenario Generation." Noteworthy is their recognition of the analytic equivalence between inside-the-head and computer-simulation generated stories about the future. Because of its sophistication, and yet also because of its failure to mobilize social science theory for its purposes, the work of this Rand team deserves attention.

These researchers computerized two intuitive logics to produce hundreds of simulated trajectories or scenarios. Drawing on informal models of the expansion of Islamist extremism and

nuclear weapons use, they programmed computers to produce, via Monte Carlo randomization, hundreds of distinctive outcomes for each model. By doing so, they claimed to map the "possibility space"—to show the range of possible outcomes that the two models forecast for the spread of Islamic extremism and the next use of nuclear weapons. Structurally, this is the path we advocate for addressing problems in Quadrant 2, where relevant theoretical ideas are available but data are scarce or non-existent. However, we also advocate a substantially more ambitious strategy that requires much better articulated theoretical content in the models informing the computer simulation as well as in the interpretation of results.

In their first exercise with Massive Scenario Generation, the authors started with a "reasonable but untested analytical model" that needed "substantial" ad hoc "enhancements" to yield stories relevant for their purposes. That model, an unpublished treatment of "extremism" as a "disease," had been developed by one of the authors (trained as a chemist), without regard to data or theory about Muslim political mobilization (Davis, Bankes, et al., 2007, p. 26). To cope with the exogenous uncertainties surrounding any potential instance of Islamic extremist mobilization, they added stochastic inputs (Davis, Bankes, et al., 2007, p. xiii). For their second application, the authors tried to develop their own model of New Nuclear Use by brainstorming a list of key situational variables. When that approach failed to produce a coherent model for computerized operationalization, the authors settled for a "first-cut" dynamic model of their creation, tinkered with its mechanisms "of influence and causality," and pruned the number of variables down to nine so that when computerized, it yielded "interesting scenarios" (Davis, Bankes, et al., 2007, pp. xiii, 44).

Despite the informal procedures and absence of explicit social science theory for developing their models, the authors present their work as evidence of the value of explicit models as the basis for computer simulations that can produce scenarios useful for forecasting. That, they emphasize, is the most promising way to systematizing the "intuitive logic" approach to scenario production. But the authors are modest in their claims to have fully illustrated this point. For their description of the New Nuclear Use model sounds very much like BOGSAT. And they concede that their list of starting-point conditions might well "have been developed by, for example, a political scientist writing a thoughtful essay or a strategic-planning exercise in which a number of experts brainstorm about the New Nuclear Use" (Davis, Bankes, et al., 2007, p. xii).

In principle, the exercises in this study were designed to establish boundaries for the possibility space. Leaving aside the acknowledged weakness of the models used, the authors offer other cautionary observations. They do not claim to distinguish degrees of probability within the spaces of the possible. Nor do they believe that the kind of structural models operationalized in these experiments, as opposed to agent-based models, are capable of serving as "the engine for non-intuitive emergent phenomena" (Davis, Bankes, et al., 2007, p. 13n). Still their work constitutes an important link between extensive, non-computerized, efforts to discipline scenario techniques for better forecasting and the theory-informed deployment of ABM computer simulation we advocate and illustrate here.

Two studies in 2007 evaluting the potential of computer simulation saw the technique as having great promis, but they both focused on computer models pitched at the generic level, not on their potential for forecasting or strategy evaluation based on virtualizations of time- and place-specific problems (Harrison et al., 2007; Davis, Eisenhardt, and Bingham, 2007). To aid policy-making, as well as validate predictions against real-world outcomes, simulation models must deploy theoretical knowledge to link forecasts to causal mechanisms operating in context. Otherwise, the model sheds little light on opportunities for mitigating threats or exploiting opportunities. For example, assessing the likelihood of genocidal violence against Muslims in India requires a theory of how India, and similar countries operate (i.e. a coherent mostly inside view). For such problems, we do not claim that social science theory comes close to the ability of physical scientists in domains such as nuclear reactions or pilot training to capture the forces and contingencies that produce success or disaster. We do assert, however, that in quadrant 2, where analysts confront rare but high impact events, the intelligence community and its support staff of think-tanks and academic experts, have largely failed to exploit available high-quality social science theory in combination with advanced computer simulation techniques to navigate the space of the plausible and guide policy makers toward better choices.

The IC's propensities to treat problems in quadrants 2 and 4 as if they were quadrant-3 problems, where data are plentiful but theory scarce, have been major reasons for this failure. This is a mistake leading to, but also caused by, overinvestment in area-specialist knowledge and, more recently, in machine-learning, techno-empiricism. This latter tendency is well-illustrated by DARPA's Integrated Crisis Early Warning System (ICEWS) project. ICEWS was initiated as a tournament that challenged performers to provide early warning of international crises anywhere in the world (both domestic and international) and arising from any possible source.[6] By one well-informed estimate, expenditures on the program from 2007 to 2015 "probably roughly equaled the whole of NSF spending on all international relations and comparative politics research" (Schrodt, 2015, March 30). Although more successful than other related efforts, ICEWS never achieved its goals and failed to yield technologies for forecasting or analysis that were transitioned into official operational use.[7] The core problem was its over-reliance on the outside-view—on attempts to generate forecasts solely from available data patterns, uninformed by the kind of inside-view that theories of international security at the global level and of domestic politics at the nation-state level could have provided.

The overarching ICEWS goal was to "to develop a comprehensive, integrated, automatic, generalizable, and validated system to monitor, assess, and forecast national, subnational, and international crises in a way that supports decisions on how to allocate resources to mitigate them" (O'Brien, 2010, p. 89). At the kickoff meeting, the competing research teams were told that: (1) the focus would be on forecasting high-impact, "negative" events;

(2) the program would exploit and extend machine learning techniques and natural language processing for analyzing immense quantities of textual data. DARPA's reliance on technology and data over theory was evident from the domination of statisticians, mathematicians, engineers, and artificial intelligence experts on both performer teams and the DARPA management team. Social scientists with relevant theoretical knowledge were in short supply.

Early on, the DARPA project manager for ICEWS, and his staff, recognized the limits of purely statistical models. Committed to causal modeling, and the verification of tools that only causal understanding can provide, they saw the need for social science theory. But they also recognized a key problem: relevant social science knowledge was not consensual. Theories abound, but social scientists do not agree on which are valid and the engineers and natural scientists in national labs and the intelligence community tend to hold skeptical, if not dismissive, views of social science. At one DARPA workshop a leading physicist made his feelings clear: "I laugh at social science theory." While interested in building computer models of human societies, he was entirely unfamiliar with theories of social, economic, and political change and so had no inkling as to how they could be used to build the models he wanted. The systematic mobilization of social science theory for forecasting is also confronted with skepticism from the opposite quarter. Many IC analysts and foreign policy experts received training in the humanities or in ideographically inflected, discursive fields of history or area studies. They tend to be as unfamiliar with, and as uncomfortable with, social science theorizing as are natural scientists.

Gaps and flaws in available theories of the social world certainly restrict the extent to which high confidence simulation models of particular domains can be built. The larger and more granular the domain space, the harder it is to build reliable, theory-informed simulation platforms. Yet, as we show below, effective simulation models of complex political problems have been built using well-developed theories of institutions, collective identity, and individual and group mobilization. However, even when theories strong enough to be useful are available, there are two more challenges to building simulation tools suitable for problems located in quadrant 2:

1. the integration or federation of theoretical models;
2. their verifiable operationalization.

Models of discrete but overlapping phenomena cannot simply be aggregated. Social scientists standardly develop their models by separating out targeted phenomena of their work from other elements "in the wild" with which they are enmeshed. A model of secessionism might, for example, include regional levels of group discontent, but omit differential mobilization of distinct discontented populations. As a result, the boundary conditions and variable definitions of trusted models may not naturally articulate with those featured in models of even closely related phenomena. That means there is real work to be done stitching theories and models together if effective simulations of a multidimensional situation are to be constructed. To work within a computer simulation and to remain faithful to the substantive logic of the underlying theories, we need to articulate models that interface appropriately. In other words, an array of putatively valid models of particular aspects of a phenomenon is not the same thing as *a* valid model of the larger system. Only by federating models developed and tested separately can the simulation target be realized effectively enough to provide an inside-view.

To take a hoary example: imagine building a simulation of an elephant without a shared integrative theory of what an elephant is. If the simulation were to be constructed from an array of separately developed models of tails, ears, tusks, feet, hide, and brain, it would be necessary to: (a) standardize the terminology and measurement units of these separate models; (b) verify that each separate model makes no assumptions or includes no key tenets that contradict key elements in other models. Of course, if the theorists possessed a prior, and powerful theory of "elephants," these problems would be surmountable. But if they had never seen an elephant, or even tried to imagine one, and had no common theory of how to link the ears that interest one modeler to the tusk of interest to another, there is little reason to expect that the individual models could be woven into a coherent integrative picture of the beast itself.

Simulation is thus both science and art: it requires using incomplete theories to integrate multiple models to capture the dynamics of a targeted socio-political space. The variety of available theories, often focused on several levels of analysis away from that of prime interest to intelligence analysts and policy makers, means that not all forms of simulation modeling can be harnessed with equal success.[8] As the ICEWS project manager recognized in his assessment of the project's strengths and weaknesses, it was only via computer-supported agent-based models capable of absorbing and integrating different kinds of causal theories that the potential of integrating social science theory and computing power could pay off for policy purposes.[9]

Unfortunately, the problem of "virtualizing" a complex policy space does not end with a platform capable of hosting and integrating models of different types. Each element must also be operationalized in ways true to the claims of theories to which it is connected. This is both a modeling and a programming problem because theoretical propositions must be articulated into formal algorithms. How difficult it is to do this will vary with the complexity of the model federation and the talents of model integrators—both those responsible for mobilizing and integrating separate models and those responsible for faithfully reproducing those models in a virtual environment.

Key to solving the operationalization problem is understanding the phenomenon of emergence, a concept at the heart of complexity theory and a key element in agent-based computational modeling. Many dynamic multi-body problems, or domains with multiple causal processes operating simultaneously, manifest emergence—a propensity to produce patterns of outcomes that themselves become mechanisms driving outcomes at a higher level of analysis. Hundreds of birds interacting as they fly produce a flock, with dynamics and patterns of behavior utterly distinct from those of individual birds. A market or a political system operates

according to laws and regularities unpredictable by and unreducible to the actions of individual consumers, producers, voters, or politicians.[10]

Any sophisticated simulation must rely on emergent properties at the societal level to yield patterns of complex behaviors and processes. As Holland has explained, emergence manifests itself in complex systems (cellular automata, board games, biological evolution, poetry, markets, international politics, etc.) in which "rule-governed entities interact" (Holland, p. 132). In such networks, elements with distinctive properties are constrained by rules that govern what interactions can take place but not the result of those interactions. The natural appearance of emergence in agent-based models—featuring, as they do, a finite range of autonomous elements regulated by algorithmic rules that limit the interactions among them--is key to what gives them such potential for translating knowledge of behaviors and circumstances into mechanisms of regularized transformation and then into probability distributions of outcomes. This is how agent-based models can capture the reciprocal interdependencies, topological variation, network specificities, agential behavior, and other complexities of most strategic and foreign policy questions.

For example, there is no gene for an elbow. But lower level genetic code and epigenetic processes include directives that produce proteins and other effects in complex combinations that reliably result in elbows. Theory-based simulations work in the same way to yield effects readable as phenomena not specifically programmed to appear, such as motivations, emotions, collective mobilization, and cultural differences. These and other patterns that arise from more basic elements are prominent variables in theories being operationalized. Normally they appear at a higher level of abstraction than the elements of the simulation model. Accordingly, simulation designers must understand how to calibrate algorithms and parameter settings so as to yield a variety of phenomena that are not themselves part of the program generating the simulations, including violence, corruption, secession, anger, alienation, loyalty, and prejudice.

4 | COMPUTER SIMULATION: GETTING WHAT WE NEED WHEN WE CAN'T GET WHAT WE WANT

Ideally, we could observe possible futures and possible pasts unfolding in response to imagined policy interventions. With due apologies to Mick Jagger, we can't always get what we want. But if we try sometimes, we just might find we can get the simulations that we need—data from worlds with substantial theoretical isomorphism to our own and tuned to an approximation of the antecedent state of affairs. The road to realizing the potential of validated and verified simulation models to address quadrant 2 problems will be long—measured in decades—but scholars have made more progress than is generally understood. Further progress will require appreciating the challenges of both underdeveloped and misapplied social science theory. It will also require translating good theory into effective simulations and an appreciation of tools that can be adapted for quadrant-2 problems, where the stakes are high, uncertainty is impossible to eliminate, and relevant data are scarce or non-existent.

Here we illustrate how the Virtual Strategic Analysis and Forecasting Tool (VSAFT), a sophisticated agent-based modeling platform for conducting disciplined thought experiments on virtualized political systems, offered support to the U.S. government in crafting policies toward Bangladesh in the summer of 2013. This was one of a variety of similar projects, conducted over the last two and a half decades, designed to map the state space of possibilities in complex policy domains. For each problem, researchers simulated possible futures and possible pasts to generate a systematic outside-view, a perspective from which the actual world that did occur, or would occur, could be treated as a sample from a large distribution of possible trajectories. Among the questions addressed by these projects were the following:[11]

1. From the late 1990s, looking forward three decades, at what rate would semi-authoritarian but USA-friendly regimes in the Middle East undergo political transformations and develop into Islamist-dominated states?
2. In 2000, what were the 5-year prospects for Pakistan to remain a turbulent, culturally mixed polity marked by recurrent violence, corruption, and challenges to its authority or to develop into an orderly democracy; or to be transformed into an Islamist-fundamentalist state?
3. In the summer of 2002, looking forward 1–3 years, what would be the probable impact on the stability and political orientation of Muslim-majority USA-friendly regimes in the Middle East under conditions of cycles of intifada-style violence between Israel and the Palestinians or of large-scale violence elsewhere in the region? How would US-sponsored diplomacy and repression by regional states impact these probabilities?
4. In May 2011, what would be the effect on violence, casualties, and insurgent strength in the Kundoz and Kandahar provinces of Afghanistan of a shift in policy toward direct support of tribal leaders?[12]
5. In January 2013, what would be the implications for the stability of Venezuela and the political complexion of its government of the death of President Hugo Chavez?
6. In 2016, with what probability could available courses of action (those considered as options but not adopted) have reduced the scale of civilian deaths in Syria between 2011 and 2014?

5 | THREATS TO BANGLADESH DEMOCRACY AND STABILITY IN 2013

National elections were scheduled in Bangladesh for late 2013 or the beginning of 2014. In the spring of 2013 both Bangladeshis and U.S. policy makers worried that the country's fragile democratic institutions could be damaged or destroyed by distrust and turmoil surrounding them. Lurking behind democratic contestation in Bangladesh were ferocious political-religious rivalries centered

round a polarizing division between the incumbent Awami League Party and the Bangladesh Nationalist Party (BNP). Against a background of paralyzing disputes in the past over rigged elections, a law was passed requiring a non-party caretaker government anchored in the civil service and the judiciary to be installed to oversee elections. Past experience in Bangladesh with such governments, unaffiliated with either the major parties or with the military, had been positive. But the Awami-dominated government decided not to implement the law. It would itself oversee the elections, and promised fairness.

The stakes of the contest were high and the BNP had good reason to doubt the Awami League's commitment to fairness. Both parties had a history of corruption, vote-rigging, and using incumbency to prosecute opposition leaders. U.S. interests in Bangladesh were focused on preserving stability, fostering consolidation of Bangladesh's fledgling democracy, and avoiding either a military takeover or an upsurge of Islamism. The U.S. experts believed that American diplomacy and especially messages sent to the Awami League government, would have a non-trivial impact on outcomes. The experts also saw three scenarios as roughly equally plausible—an assessment grounded in an inside-view of Bangladesh:

1. A caretaker government is given power prior to the election period.
2. The Awami League maintains control of the government prior to and during the election.
3. A military coup occurs in the period just before the election.

VSAFT's assignment was to assess the relative likelihood of these outcomes in the absence of U.S. political and diplomatic intervention, forecast their implications for political stability, and provide guidance to policymakers confronting choices among various options for promoting a democratic and stable Bangladesh.

Ian Lustick and collaborators originally developed a computational (agent-based) model of Bangladesh as part of the Lockheed Martin ATL's contribution to the ICEWS tournament. Built originally in 2008, it was refined and deployed on a quarterly and then monthly basis for 7 years, generating, each time, a batch of 1,000 unique year-long trajectories. By 2013, Virtual Bangladesh was one of more than fifteen country models generating monthly forecasts that formed a substantial basis for assessing the accuracy, precision, and recall of forecasts of standard "events of interest."[13] To assess the likelihood and effects of a caretaker government installed prior to the election period (Scenario 1), Virtual Bangladesh was used in May 2013 to produce 1,000 simulation trajectories of Bangladeshi politics ending in May 2014. The projected election period of November-December was subjected to punctuations of higher than normal turbulence and elevated political distrust. The year-long forecast window included 5 months of post-election forecasts to see how different scenarios varied with respect to prospects for stability in that period. The distribution of outcomes in these trajectories formed the basis, after each deployment, of forecasts and analyses of political futures for the twelve subsequent months.

5.1 | Scenario 1: A Caretaker Government

Futures featuring a caretaker government were explored by focusing on the subset of model trajectories that developed to include coalitions directed by the bureaucratic state apparatus, the civil service, and judiciary. The spine plot in Figure 1 maps the state space of the future of Bangladesh from November 2013 through May 2014 along two dimensions. Along the x-axis are the political groups dominating the governing coalition. The width of each segment with a group label indicates the proportion of 28,000 simulated weeks (1,000 seven-month trajectories) during which the sum of the influence exercised by the group's agents exceeded that of any other group. On the y-axis those segments are color-coded to indicate the proportions, within each group's periods of dominance, of degrees of political stability. "Calm," marked by green, indicates an absence of significant violence or protest. Intense Politics, marked by blue, indicates political agitation or violence but not both. "Instability, marked by yellow, features widespread protest or violent unrest. "Severe Instability," marked by red, indicates the highest levels of both protest and violence.[14] The width of the columns yields information about how often the group represented by that column enjoyed an aggregate influence greater than that of any other group. The coloration of the column—amounts of green, blue, yellow, and red—indicates how stable or unstable were political conditions in the weeks of the group's dominance.

Our strategy was to look at the 17% of the event space within 1,000 trajectories featuring domination by "the state," that is, by the organizational apparatus of governance itself—anchored in the civil service and judiciary but excluding the military. This is the column in Figure 1, labeled "State." A noteworthy feature of periods dominated by caretaker "State" dominated governments is that they include more "Calm" (more green) than any other column, although they also include more "Severe Instability" than any other, and nearly the maximum proportion of Severe Instability as compared to the periods of domination by other types of Bangladesh governments.

Figure 2 zooms in to examine only the distribution of outcomes within the "State" dominance column in Figure 1. Here we can discern which types of state-dominated governments formed, and which coalitions were in power during periods of stability or instability. Along the x-axis in Figure 2 appear the most significant coalition allies for a caretaker government. When a group is in coalition with the dominant group, the model treats it as an "Incumbent Group." The width of the columns indicates how often that group appears as the largest coalition ally. Again, the height of the red, yellow, blue, and green bars within each column indicates how often coalition alliances, between state-run caretaker governments and another major group in Bangladesh, are associated with each stability level. The display shows, for example, that it was rare indeed for Virtual Bangladesh trajectories featuring dominance by the state apparatus to exhibit "Calm" politics (green) unless the state drew support from significant coalition allies (incumbents). (Note the very narrow

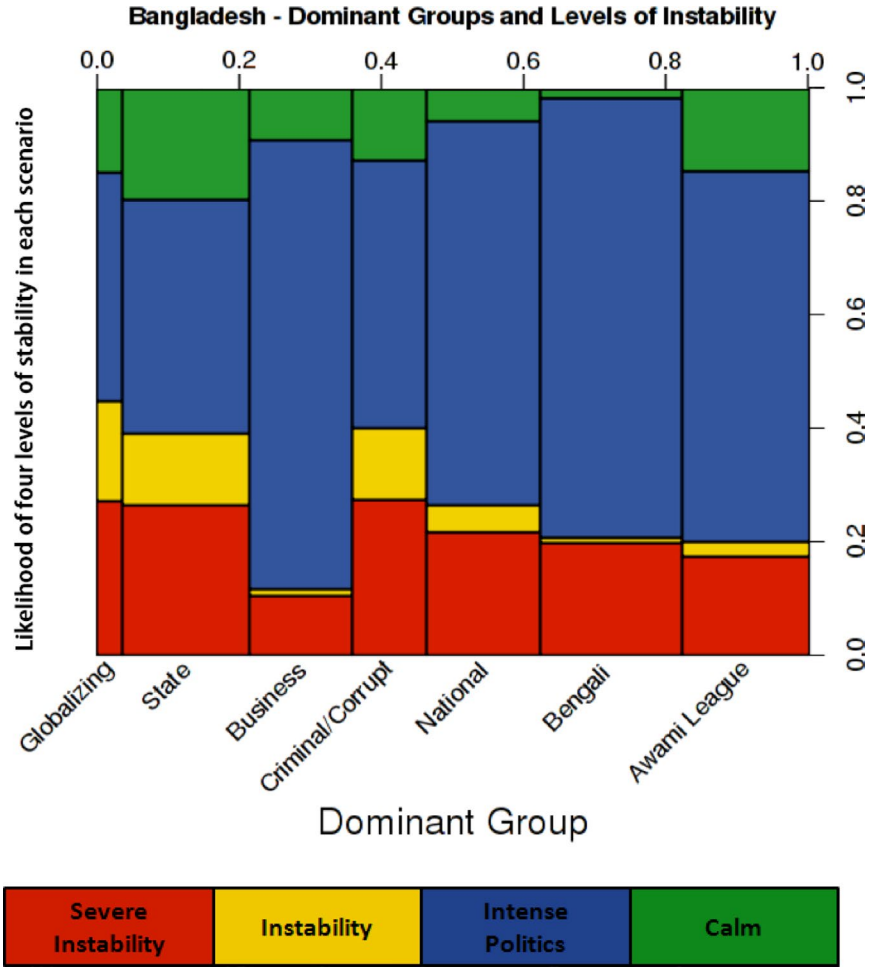

FIGURE 1 Baseline results, with caretaker government subset highlighted

band of green along the top of the column labeled "No Significant Incumbents" in Figure 2.)

Specifically, if a state-run caretaker government can draw substantial support from the business community, from Western-oriented globalizing interests, from broad Bangladeshi nationalists, or even from corrupt elements of society who want the caretaker government to succeed, the prospects for a stable election period are relatively good. The model puts a probability of about 35% on one of these coalitions developing. (The combined widths of the columns in Figure 2 apart from the column labelled "No Significant Incumbents.) Within that zone of the possibility space, there is a small chance of Instability or Severe Instability, with the remainder of the forecast fairly evenly split between Calm and Intense Politics. Notably, if the caretaker government can use a broad-based nationalist appeal to anchor its coalition, the risk of Instability and Severe Instability disappears, and the likelihood of a calm election and aftermath is over 50%. (See the narrow column in Figure 2 labeled "National.") In the 65% of the distribution, where the caretaker government cannot find a potent coalition partner, the likelihood of Instability or Severe Instability is nearly 60% while that of Calm is tiny (see the wide column in Figure 2 labeled "No Significant Incumbents").

5.2 | Scenario 2: Awami League Government Oversees Elections

Refusal by the Awami-League-dominated government to dissolve itself would result in elections conducted under its supervision. We see from the width of the far-right column of Figure 1, labeled "Awami League," that this condition was true in approximately 17% of the simulated space of Bangladesh's future from June 2013 through May 2014 (see the width of the column labeled "Awami League" in Figure 1).[15] Compared to the column associated with rule by a caretaker government (labeled "State") this zone of the space of possible futures features a larger proportion of weeks coded as Intense Politics and smaller proportions of Calm, Instability, and "Severe Instability." To explore the implications of an Awami League government that maintained power through the planned elections we consider, in Figure 3, only the portion of the forecasted distribution of outcomes where the Awami League is the dominant political force. Listed on the x-axis are the Awami League's primary coalition partners across the weeks of its dominance. To be clear, Figure 3 is simply a disaggregation of the column labeled "Awami League" in Figure 1).

Here we see that Awami League governments, with allies drawing on bureaucratic state institutions or on Bengali

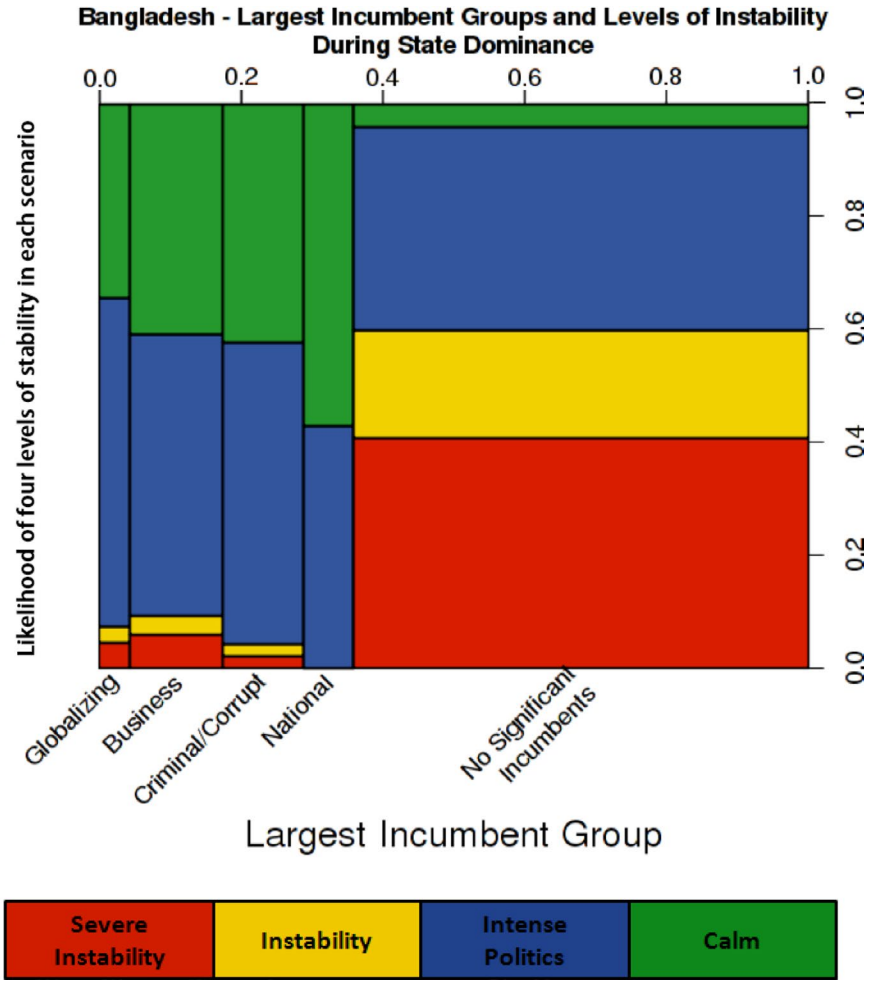

FIGURE 2 Coalition partners under a caretaker government

ethnonationalism (Bengali) or Bangladeshi patriotic nationalism (National), were likely to navigate the election period with no risk of major instability. However, in more than 80% of the weeks featuring an Awami League government (the broad column labeled "No Significant Incumbents" in Figure 3), these allies were absent. Much more probable, given Awami League dominance, were partisan and relatively isolated governments. Here the odds of severe instability rose markedly while prospects for Calm fell sharply (note the broad band of red and the very narrow band of green in the "No Significant Incumbent" column compared with the absence of red elsewhere in Figure 3). Nonetheless, prospects for avoiding instability were significantly better under narrow Awami League governments than under caretaker governments lacking substantial political alliances.[16]

5.3 | Scenario 3: Military Takeover Prior to Election

The political dominance of the military in Bangladesh occurred very rarely in the 1,000 simulation runs for the period in question, hence its absence as a column in Figure 1. But in 2013 U.S. observers looked back on a history of military intervention in Bangladesh politics. In the face of a seemingly intractable divide between the Awami League and BNP, and after a decade of work stoppages, protests, violence, and chronic corruption, they understandably wondered whether the military would intervene to stabilize the country, as it had in 2007–08, and whether it would succeed.

Because sudden seizure of power by the military did not emerge in the space of simulations traced by our 1,000 trajectories, the implications of this scenario were explored by doing a "what-if" experiment: rerunning the simulation but inserting into each trajectory an attempted military coup on November 1 causing the cancellation of the elections. We tested two experimental conditions to simulate different types of coup attempts, each with a precedent in the history of Bangladesh.

1. High-level coup: The upper echelons of the military seize major governing apparatuses and assert their authority, particularly in the national capital, Dhaka.
2. Low-level coup: Lower-ranking military officials seize local governing institutions.

Both "punctuations" explored the scenario of military intervention in politics, but not necessarily military governance. In these what-if scenarios of military attempts to depose the Awami League, whether the military succeeds and whether it then

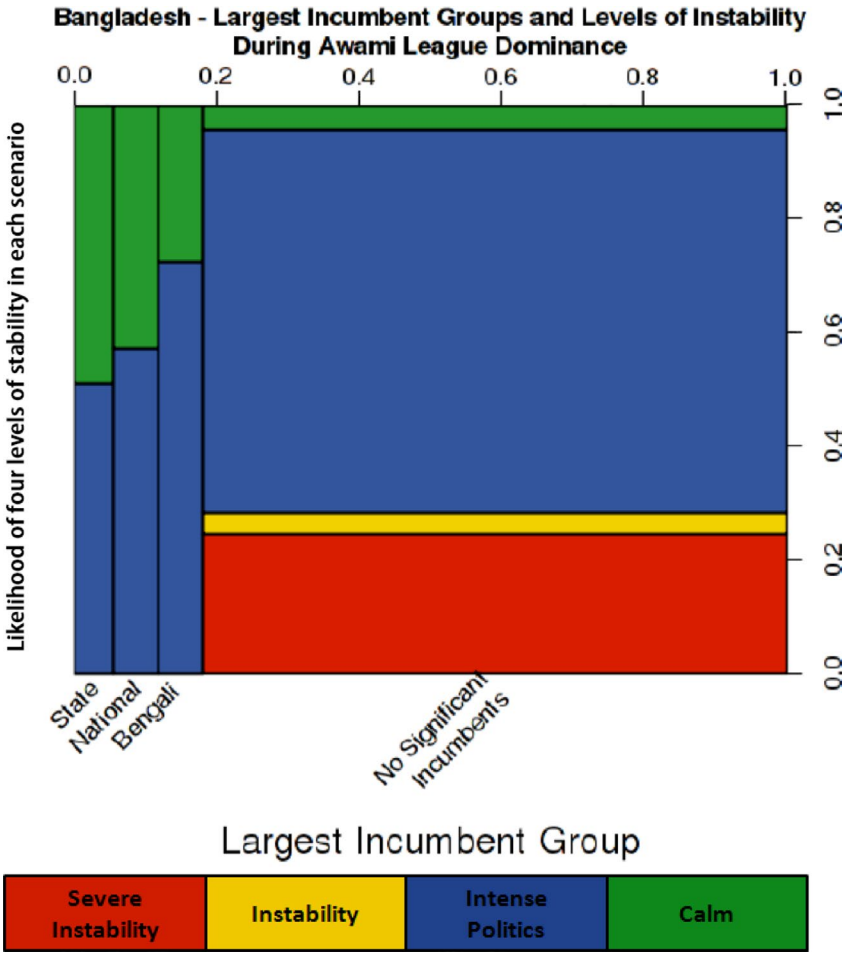

FIGURE 3 Coalition partners in Awami League governments

remains in power to govern, or instead steps aside to make room for another political actor, are empirical questions answerable from model output.

5.3.1 | High-level Coup

Figure 4 displays data from the first version of a military-takeover scenario—a high-level coup—and pertains only to November 2013-May 2014. As before, the x-axis shows likelihood of political dominance by group, and the y-axis shows stability patterns associated with dominance by each group. Given a High-level military coup, there is a significant chance of governments led by the military, but also by Business or Bengali nationalist elements, or a return of the Awami League (indicated by widths of columns in Figure 4).[17]

5.3.2 | Low-level Coup

Figure 5 displays the results from the Low-level coup scenario. The distribution of likely dominant groups and their stability patterns resembles the outcomes observed in the aftermath of a High-level coup. Figures 4 and 5 show that in both conditions the probability of Instability or Severe Instability (the yellow and red portions of the display) is high during military rule after either kind of coup and increases by about 20% from the High-level to Low-level coup condition (comparing the amount of red and yellow in the columns labelled "Military" in Figures 4 and 5).

6 | POLICY IMPLICATIONS OF SIMULATION RESULTS

Implementing a caretaker government for diffusing the contentious pre-election rivalry between the Awami League and the BNP was a viable option for Bangladesh, but by no means a risk-free solution. Without other intervening factors (e.g. different U.S. policies toward Bangladesh) the caretaker government had less than an approximately 40% chance of heading a government confronting unrest and instability (red and yellow areas as a proportion of the entire spineplot displayed in Figure 2).

Relative to the caretaker-government scenario, Awami League government management of the election was a lower-risk, lower-reward option, registered by amount of red within the Awami League column in Figure 1). The lower risk of Instability or Severe Instability was nonetheless linked to a lower likelihood of a Calm election period and government transition. marked by the extremely narrow band of green within the Awami League Column in Figure 1. As indicated in

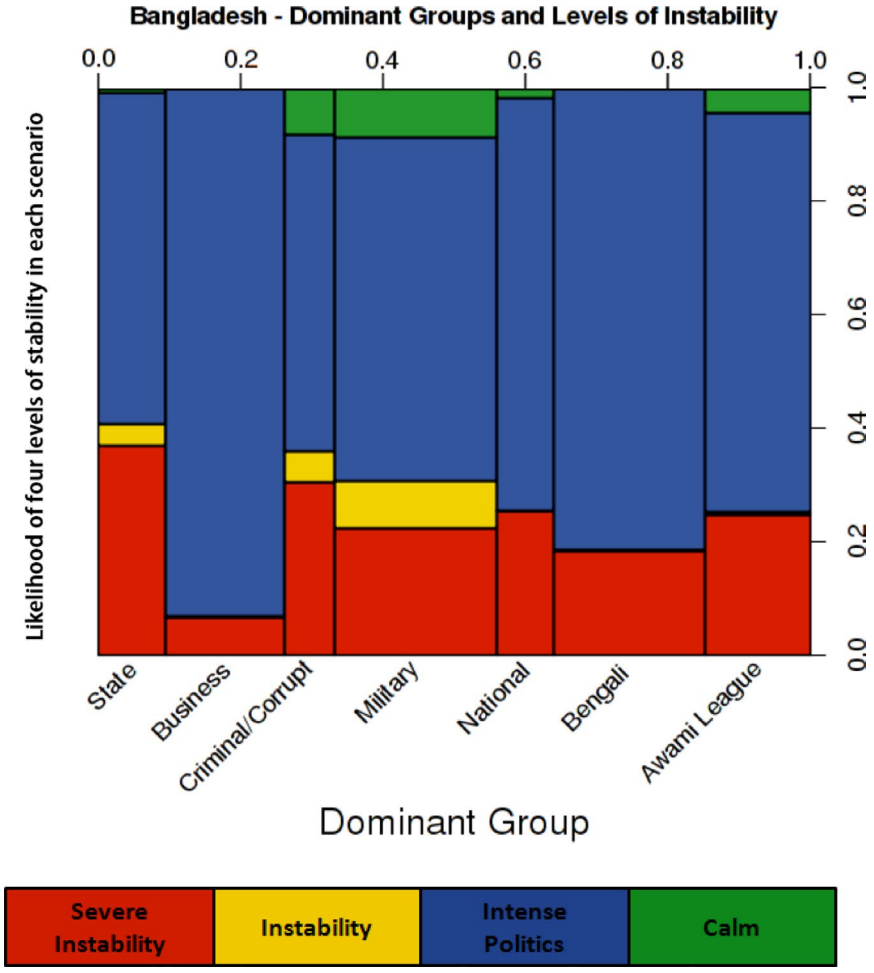

FIGURE 4 Experimental "what-if" results for a High-level coup

Figure 3, the chance of Calm increased if the Awami League broadened its coalition to include major state or ethno-nationalist actors. However, its probability decreased sharply under Awami League rule, without significant coalition partners, and this did indeed register as the much likelier outcome of military intervention.

These results suggested that military intervention attempts would not substantially alter the probable political complexions of emerging governments. However, both High- and Low-level coups did produce an approximately 21% likelihood of a military-dominated government (the widths of the columns labeled "Military" in Figures 4 and 5)—an outcome that did not appear in the baseline results. They also reduced the likelihood of a caretaker government, reflected as a narrowing of the "State" bar (farthest left) in Figures 4 and 5.

In 2013 the United States, Bangladesh's largest trading partner, had a strong interest in the country continuing its role as an anchor of American diplomatic and security interests in South Asia and the Indian Ocean and as an important asset in the worldwide counter-terrorism campaign. Bangladesh served as a model of a politically moderate Muslim majority with a democratic regime and a capable military deployed regularly on United Nations peace-keeping missions. Political corruption and violence associated with intense Awami League/BNP competition threatened these interests. In that context, Washington wanted well-run and fair elections, but it also wanted stability.

In this light, simulation results suggested that the best possible trajectory would be a caretaker government with enough broad political support to resist interference by the two large partisan contenders. To the extent the United States was in a position to foster nationalist, patriotic, civic, or business support for "clean" elections and to the extent that the caretaker government was seen as staffed by officials not beholden to either the Awami League or BNP, that course of action would yield the best outcome.

However, given that 65% of the futures featuring a caretaker government did not include broad support for it (the width of the "No Significant Incumbents" column in Figure 3), a decision to push in this direction was characterized as risk-acceptant because of the high probability of instability, even severe instability. The safer option identified was to coax the Awami League government to broaden its political base, send credible signals of a shift in its partisan orientation, and/or share real power with rivals in the period prior to the election. While success in this effort would not guarantee the absence of instability or fair elections, it would reduce the probability ratio of undesirable to desirable outcomes. The what-if military-coup simulations suggested that much depended on particulars of that intervention that would be difficult, if not impossible, to anticipate. In that light, and in light of the weakening of Bangladeshi institutions associated with a military coup and the very high risk of tumult in its immediate aftermath,

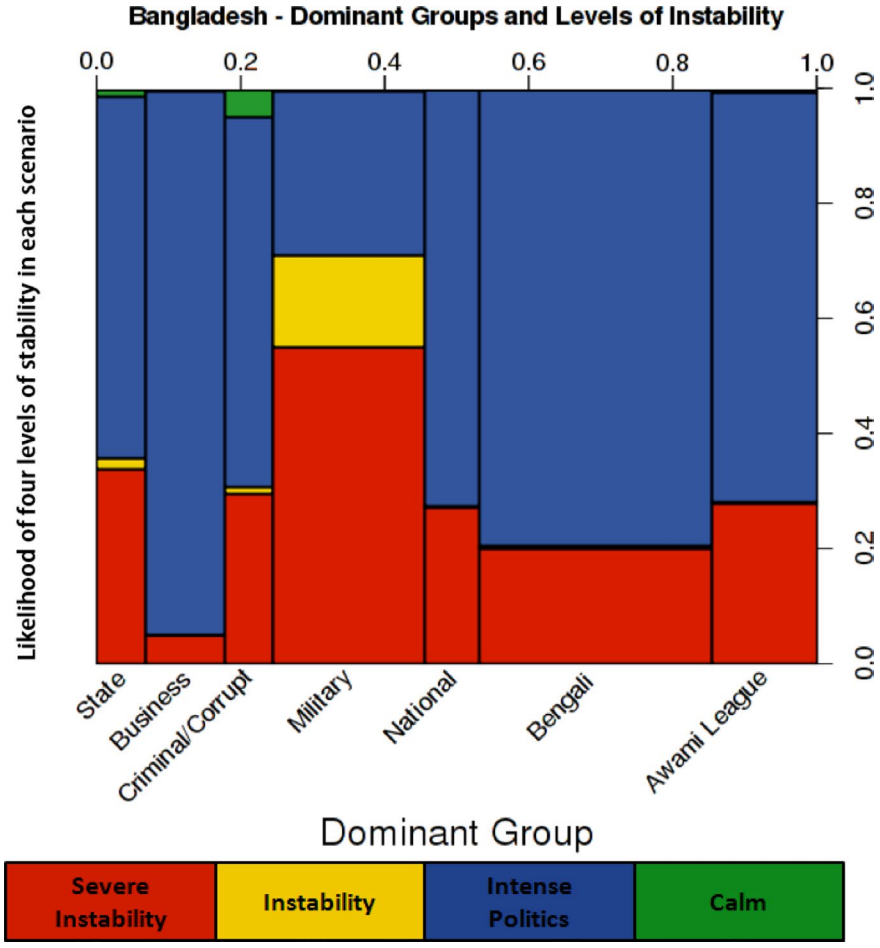

FIGURE 5 Experimental "what-if" results for a low-level coup

encouraging military intervention was characterized as a high risk and low reward course of action.

In the world as it actually unfolded, the Awami League did supervise elections directly. The BNP retaliated by boycotting the elections, which, although held on January 4, 2014, attracted fewer than a quarter of eligible voters and were marred by violence, both by opposition elements and the government. Yet by the end of the year, the Awami League managed to broaden its coalition by attracting support from Jatiya, a military and business-friendly Islamically oriented, conservative party, and inaugurated a 6-year period of Bangladeshi economic growth and Awami dominance (including prosecution of BNP leaders). In other words, the thread of Bangladesh's actual future, beginning in 2013, led through the portions of its possibility space demarcated in the rightmost columns in Figures 1 and 2, meaning that in the space of the possible mapped by the simulation model, the actual future can be located within one of the more accessible parts of that space, i.e. those containing outcomes the model considered relatively probable.

From all indications, U.S. policy makers were convinced that the Awami League would not allow a caretaker government, that elections under these circumstances would be accompanied by turbulent but manageable disruption, and that this would result in continued Awami League control. No public warnings were issued by the U.S. government prior to the elections. No public rebukes were offered afterward in response to widely reported abuses against opposition leaders and activists. In May 2014 Washington congratulated the new Awami League government on its performance, and hailed the continuation of annual bilateral strategic partnership meetings. Accordingly, although the impact of VSAFT's simulation results on U.S. messages to the Bangladesh government prior to the election is not known, we can say that U.S. assessments of what was prudent and achievable in Bangladesh, and the policies adopted in light of those judgments, were closely aligned with the findings and conclusions reported here (Office of the Spokesperson, 2014, October 29; Riaz, 2014; Vaughn, 2010).

7 | THEORIES AND OPERATIONALIZATIONS IN THE VSAFT META-MODEL

The successful application of Virtual Bangladesh to a real-world, real-time, foreign policy problem is good reason to grant some face validity to the approach; all the more so because the model was not a one-off design, trained specifically for Bangladesh, but the semi-automatic product of a general country-level modeling platform and the protocols for simulation exercises performed with it. Its success thus warrants a more affirmative verification of the model. We see

validity here as a test of whether the model makes correct forecasts whereas verification is a gauge of whether the model's construction and operation faithfully represent condensed knowledge, yielding an understanding of why outcomes occurred that enables evaluation of strategies for goal achievement. We thus now turn to consider briefly how substantive theories were deployed, integrated, and operationalized to produce model effects.

As noted above, the Bangladesh simulation exercise was one of many that used VSAFT models or precursors. For each application, simulations yield a statistical summation of possible worlds, an outside view, to analysts who lacked systematic data but had plenty of inside-view hunches about the causal propensities of specific entities. The extent to which analysts treated model results as valid reflected the extent to which they, as South Asia experts, were willing to treat Virtual Bangladesh as if it were the country itself—and the weight analysts gave to the track records of similar VSAFT models on comparable forecasting tasks for other countries.[18] Treating the simulation as "verified," meant accepting the faithful integration and operationalization of valid social science theories (Lustick & Tubin, 2012). In the following section, we illustrate verification by explicating which theories are deployed within the meta-model of which Virtual Bangladesh is a particular instantiation, and how they are linked to one another and operationalized.

The VSAFT meta-model is a template for producing virtualized political arenas—and has produced dozens of country virtualizations, including Yemen, Iran, Egypt, the Philippines, Malaysia, Vietnam, Venezuela, Indonesia, and Thailand. At its most fundamental level, the meta-model was designed to correspond to a national society governed, weakly or strongly, by a central state and organized around a common arena of direct and indirect interaction. Depending on the size and complexity of the target country, a VSAFT model is populated by two to seven thousand boundedly rational actors whose networked interactions yield a complex adaptive system impacted by a mix of predictable and unpredictable fluctuations on dimensions relevant to actors. Agents, imaginable as differentially influential individuals, households, localities, or positions, plus attributes and identities (affiliations), are the constituent elements of the system. Sustained interactions at the micro-levels produce patterns of agent behavior and identity complexion. In accordance with the principle of emergence, these patterns become mechanisms at "higher" ontological levels. For example, larger repertoires of available affiliations among low echelon influential agents can produce bureaucracies more responsive to the populations they serve (or more corrupt), and less likely to serve as effective transmission belts for the preferences and affiliations of higher authorities.

The epistemological foundation of the VSAFT meta-model, tunable with real-world data to produce political representations of particular societies at particular moments, conforms to the reigning paradigm of contemporary social science within which the overwhelming majority of political scientists, sociologists, and economists work. Individuals and groups with varying influence pursue interests bounded by cognitive heuristics, institutional constraints, socio-economic inequalities, cultural dispositions, and historical legacies. Within the models produced on this foundation, the formation and transformation of elites and masses are dynamically determined by competitive and cooperative interactions. Political systems are understood as institutionalized relationships among calculated and uncalculated processes of resource allocation and competition. Individuals participate in groups that serve as communities of trust for those affiliated with them. With some exceptions, memberships in such groups are not mutually exclusive. Affiliations change with circumstances and in response to boundedly rational decision making. As associations within and between groups expand and deepen, stability is enhanced. Instability is associated with alienation and fragmentation.

Two computer modules combine to form the VSAFT meta-model. The General Political Module (GPM) is comprised of individual agents with varying influence, networks among these influential elites, and emergent groups. It hosts a second module, the Dynamic Political Hierarchy (DPH). The DPH translates changing patterns of overlapping or cross-cutting affiliations into classifications of each group as standing in supportive or oppositional relations to the dominant group, thereby affecting how each group's discontented members mobilize (legally, semi-legally, or violently).[19] It is particularly important that decisions on which theories to use, how to express them, and how to translate outputs of each theory into inputs for the others, are all made, within these two general models, prior to VSAFT's virtualization of any particular country. This procedure provides a natural demonstration of VSAFT's robustness across very different cultures, time periods, and political systems. In this way, the process is also protected against cherry-picking that could compromise the integrity of the modeling process by tuning individual country models to conform to modelers' hunches about variation across country targets.

The ontology and operational codes for the GPM and the DPH reflect and were guided by four core social-science theories: constructivist identity theory, state-society relations theory, cleavage theory, and the theory of nested institutions. As emphasized above, mobilizing effective theory to construct simulation models is the essential first step toward their conceptual integration and effective operationalization. To illustrate, we describe these four theories and how they are integrated and operationalized in a VSAFT country model.

Constructivist identity theory posits that individuals as well as groups embrace repertoires of changeable affiliations or identities (Aronoff, 1998; Brass & Richard, 1980; Chandra, 2012; Kowert & Legro, 1996; Laitin, 1998; Nagel, 1994; Posner, 2005). Depending on circumstances, selection processes, or boundedly rational decisions, people can shed affiliations or identities, and add new ones to repertoires, but such "substitution" processes occur less easily than "rotation" processes by which individuals or groups publicly identify with one of the affiliations or identities already in their repertoire. For instance, at the group level, a group with a shared identity or affiliation can decide in response to changing circumstances to oppose rather than support the government—or even to oppose the state (political system) itself.

The theory also posits that individuals can adopt a multiplicity of identities and operationalizes the notion with agent "repertoires" that grant agents the flexibility to "subscribe" to multiple latent identities, while simultaneously publicizing or "activating on" a single identity at any given time. The algorithms governing activation and subscription, as well as rotation between activated and subscribed identities, mimic the boundedly rational decisions governing identity affiliation in human actors. Critically, the GPM captures the boundedness of rationality by assigning a "sight radius" parameter that localizes the information available to each agent. This ability to adopt, discard, and reorganize affiliations is essential to the emergence of changing patterns of interaction within and between groups and thus to collective action.

Individuals and groups modeled in line with constructivist identity theory operate within an institutional context guided by state-society relations theory, which treats contemporary nation-states as composed of social-organizational and artificial organizational ties (Callaghy, 1984; Evans et al., 1985; Migdal, 2001; Nettl, 1968; Tilly, 1975). Social organization refers to cultural norms, economic patterns of production, social formations, resource patterns, and kinship groups, to the extent they can be separated from artificial organization. Artificial organization refers to authority structures designed to make and enforce resource allocation decisions to the extent these structures can be separated from social organization. Treating state and society as interdependent but distinguishable, the theory features networks of "elite" agents, possessing distinctive attributes, that influence and are influenced by both their elite and "basic" neighbors. The degree to which these networks are "embedded in" or autonomous from society critically influences how societal agents respond to the state.

Power relations within societies are realized via cleavage theory, which posits a society comprised of individuals with multiple affiliations (Coser, 1956; Dahl, 1961; Dunning & Harrison, 2010; Lipset, 1960). Groups consist of individuals sharing an affiliation. Trust of individuals who share an affiliation, that is, within-group trust, is higher than trust among individuals who do not share an affiliation. A cross-cutting cleavage arises whenever two sub-groups within a group share a third affiliation. Cleavage theory suggests that the more cross-cutting cleavages and the higher the proportion of individuals belonging to such boundary-spanning groups, the more integrated the society and the more stable and less prone to illegality or violence that society's politics.

Cross-cutting cleavages appear in our model when two groups of agents, each defined by a different activated identity, share a mutually subscribed identity. The "repertoire" model component of constructivist theory, as described above, is critical for allowing this possibility. Just as a highly integrated society is envisioned as one in which cleavages are cross-cutting and not congruent, an integrated society in the model is a landscape in which identities are frequently shared. Over time, agents respond to circumstances not only by changing their activated identities (rotation), but by dropping an identity from their repertoire and substituting a promising alternative. By endogenizing processes of change in the social-political distance between specific groups, changing cleavage patterns across agent populations have important implications for agent mobilization. Groups closer to the dominant group via first-, second-, or third-order attachments are progressively less likely to express dissatisfaction by protest or violence.

Competition among groups, expressed within institutional arenas, is structured by changing expectations about the loyalty of protagonists to allies and legal norms. Nested-institutions theory treats the structure of political authority as a layered array of modes of competition for power based on the scope of groups' demands and expectations of compliance with norms and formal legal rules. Among political scientists, this tradition's most influential proponent in Easton (1965). His vision of institutionalization divides society into three levels: those who accept the political authority of the community, but not the legal order established by the regime; those who accept the regime's authority, but not that of the government grounded in the regime's laws and those who accept the government's authority. A fourth level consists of those in society who do not recognize, even in principle, the authority of the community (Easton, 1965; Lustick, 1993; North, 1981; Tsebelis, 1990).

This layered array of relative loyalties to the governing coalition is the core principle of the DPH, which at every time-step of the model automatically categorizes agent groups into one of five levels based on relative distance from the center of power and influence (Dominant, Incumbent, Regime, System, and Non-system). The location of a group within this hierarchy registers expectations of compliance by agents affiliated with that group with norms and legally promulgated rules. DPH levels of groups with which an agent is affiliated determine eligibility for mobilizing in various ways, ranging in intensity from lobbying to protest to violent attacks.

8 | CONCLUSION

The results of the Bangladesh virtualization exercise show that theory-informed computer simulations can far out perform handcrafted scenario-building as measured by the detail and reliability with which outcomes, causal processes, and the space of the possible, the plausible, and the probable can be mapped.

A tricker question, however, concerns the impact on judgmental accuracy. Our working hypothesis is that the advantage of VSAFT models over human intuition will extend to superior forecasting performance. But this contest has yet to be run. And it is also an open question whether VSAFT models have blindspots that the best human forecasters—for example, "superforecasters" (Tetlock & Gardner, 2015)—might be skilled at spotting. The top performers in such tournaments might well be human-machine hybrids, at least until VSAFT programmers figure out how to blend the best human insights into their models and the process can be repeated.

And an even trickier question is the impact on decision making. Although we have cited as corroboration the convergence of U.S. diplomacy with the policy inferences drawn from the exercise, we cannot conclude that U.S. policy would have been different or less

effective had the simulation results not been available. Answering the "would policy have been different?" question requires access to classified information about why key players made the decisions they did on U.S. policy toward Bangladesh during the run-up to its 2014 elections. And answering "would policy have been better?" question requires systematic analysis of the counterfactual pathways that Bangladesh's history would have taken if the United States had chosen different options.

Some might argue this is a bridge too far given the state of the art and science. They may be right but one never knows what is possible until one makes a serious effort and the IC does have sufficient experience with ABM approaches to forecasting and analysis to begin assessing their contribution (Lustick, 2016). Consider one of the most ambitious attempts to pit simulation modeling against other social-science approaches in a controversial policy debate: Would alternative U.S policies toward the Syrian conflict, 2011 to 2016, have reduced or increased the level of civilian casualties, with what probabilities? It is easy to imagine which positions critics and defenders of the Obama administration will take here. But the sponsor of the multi-method study, the Simon-Skjodt Center in the Holocaust Memorial Museum, was determined to advance beyond partisan squabbling. The Center commissioned seven separate studies, each relying on different modeling strategies and methodological techniques, including game theory, expert panel surveys, agent-based models, elite-interviewing, and configurative comparative analysis.[20] Considerable agreement emerged across studies that no plausible option available to the United States had more than a low probability of reducing overall atrocity levels, though different options would have changed the populations most gravely affected.[21]

We see great value in this type of multi-method approach to quadrant 2 problems, where of course counterfactual analysis of the Syrian civil war is located (little data; many competing theories). We also see value in more systematic testing of VSAFT techniques for theory-stiching and model construction. When do experts from different schools of thought converge or diverge in how to accomplish these critical tasks? This too will not be easy. In principle, computer simulations are fully transparent. In practice, though, any model capable of simulating political systems with sufficient nuance to help policy makers must consist of a dense array of algorithmized propositions, greatly complicating the elicitation of expert judgment.

Establishing the internal validity of agent-based models is also challenged by the emergent processes that, as emphasized, give this type of "bottom-up" modeling its power and versatility. Absent a detailed theory of emergence, efforts to trace causation will inevitably be incomplete. For these reasons, we do not expect simulations ever to reduce all quadrant 2 problems to quadrant 1 exercises in automatic calculation and deductive inference.

Nevertheless, theory-guided computer simulations can greatly assist analysts facing the daunting problems in quadrant 2. By appreciating how substantive knowledge—nomological and case-specific—can be leveraged by computer simulations that enable disciplined thought experiments, we can better resist both the siren song of quadrant 3 techno-empiricism, with its false promises of atheoretic machine-learning solutions, and the seductiveness of speculative quadrant 2 scenarios.

Of course, cautioning against techno-empiricism does not mean abjuring either case-specific knowledge or big data Deep domain expertise will remain crucial for informing models that frame the quadrant-2 challenges. That our world is swimming in digitized information is also a great resource. It makes perfect sense to exploit artificial intelligence, sophisticated statistical tools, and natural language processing for tasks such as tracing trends in sentiment, political discourse, behavior, and institutional stability. Indeed, few social science enterprises are likely to generate more data than systematic, computerized exploration of possible worlds. As experimenting with histories (traced forward in time from the present or the past) becomes a standard approach to quadrant 2 problems, technologies for interrogating and visualizing big data will be increasingly valuable supports for coping with the resulting firehose of simulation data.

We hope that reluctant intelligence analysts and academics will explore theory-guided computer simulations but we recognize many will balk at treating simulated phenomena as evidence bearing on real-world outcomes. Here we offer two suggestions for reducing the resistance, neither easy to implement. The first is to rethink the epistemic status of simulation reruns of history—and view them as theory-grounded hypotheses, with the same randomness that arises when we place confidence intervals around real-world forecasting. The second is to get in the habit of conducting validation and verification studies centered around probability distributions, not point predictions.

Both suggestions ask people to perform unnatural cognitive acts but not as unnatural as it may first appear. Remember that even traditional historical scholarship relies on sorting and analyzing large numbers of clashing accounts to come to conclusions about causal pathways and counterfactual possibilities (Ferguson, 1997; Lustick, 1996; Scheidel, 2019; Tetlock et al., 2016). Although seldom constrained to explain the theories implicit in their stories, historians do routinely conduct mental simulations of what might or could have happened that advance or rebut causal claims just as computer simulations do. Historians and qualitative social scientists should therefore view simulation as a technique that simply formalizes knowledge acquisition strategies that they have long employed. The same point applies to natural scientists, who use simulation when experimental manipulation of the real world is impossible or too risky or expensive, such as aerodynamics, ecology, nuclear weapons testing, astrophysics, and medicine. From this standpoint, it ceases to be so far-fetched that computer simulation, wedded to high-quality theory, will gradually become a standard part of the cognitive toolkit of intelligence analysts. These analysts should become accustomed to treat virtual data as "data" or, put another way, to treat all data as simulated data. This idea is not new. It is just an unusual way of stating a standard practice in science. Consider how normal it is in science for advances in technology to transform purely theoretical hypotheses about unobservable entities into data. Thus, did telescopes and microscopes,

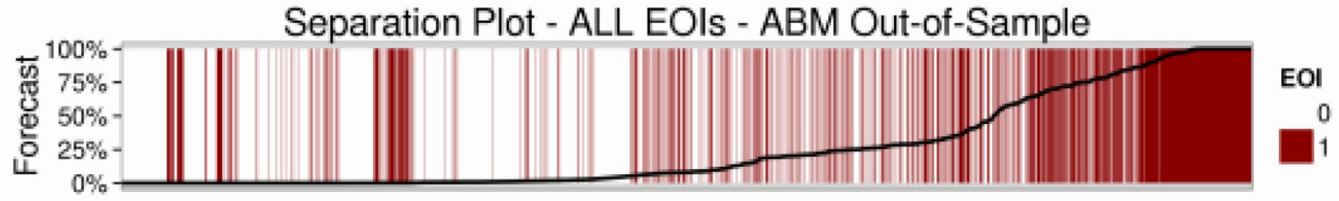

FIGURE 6 Separation Plot Display of Accuracy of VSAFT Forecasts

based on the theory of optics, transform the images (simulacra) they produced into facts (data) with which to assess other theories about stars and galaxies, or microbes and chromosomes.

Once we look at the output of simulations in this way, we can better appreciate the importance of the disciplined production process. If historians (looking backward), scenario builders (looking forward), and simulation deployers (looking in either direction) face identical epistemic challenges, the vast differences in clarity and rigor become apparent. No matter how many accounts a historian considers or scenarios an analyst generates, they have no basis for systematically assessing the probability of these stories or scenarios from the slim thread the actual world traces through the vast space of possible worlds. Nor can they identify the theories animating those accounts or verify that even if the outcome they retrodict or predict occurs, that it came about due to particular causes. This inability constrains the credibility of judgments of how a past outcome could have been prevented or of how a future outcome might be. Right now, we do not yet know how credible simulation-based retrodictions or forecasts can be, or how reliable the policy recommendations might be that flow from the systematic process-tracing that simulation enables. Nevertheless, it is the great merit of theory-guided simulation that these crucial questions can be opened for investigation. The experience of VSAFT establishes the plausibility of affirmative answers to both.

Although we see good grounds for expecting that theory-informed simulation will confer new leverage over quadrant-2 problems, there is no substitute for rigorous validation and verification studies, as recognized by the U.S. Department of Defense's formal requirement that any new technology is subjected to "Verification, Validation, and Accreditation" (VVA) before it is deployed. This procedure applies to forecasting and analysis models as well as to "kinetic" systems"(Defense Modeling & Simulation Coordination Office, 2020; Lustick & Tubin, 2012). Simulation often plays a crucial role here because any system designed for use under battle conditions cannot be tested except by setting up surrogates or simulations of the real thing. Simulations in intelligence analysis should be subjected to the same validation and verification tests, even if doing so will present distinctive challenges.

So how can we validate a technique that offers only probabilistic geopolitical forecasts? How can it be shown to be wrong, if its predictions are hedged with the escape clause that they may not conform to the outcomes observed in the one real "run" of our world to which we have access? Validation can only be accomplished by harvesting large numbers of forecasts, from one model or from many, and comparing patterns of outcomes in the real world with those produced by the model(s). That was the validation technique in the ICEWS program, which required models to make dozens, even hundreds, of forecasts each month for countries across the globe.

A "separation plot" displays the outcomes of these validation studies. Developed by Michael Ward, separation plots depict forecasted probabilities of large numbers of events against the rising curve of actual event occurrence in subsequently collected data (Greenhill et al., 2011). Figure 6 is a separation plot for a September 2015 validation study of seventeen VSAFT country models on a variety of events over several years (Lustick Consulting, 2015). Each red column represents the occurrence/non-occurrence of events coded dichotomously. Events on the x-axis are arranged by frequency. The rising black line indicates, on the y-axis, the forecasted probability of each event. A high-performing model should register a reddening trend from left to right so that events with high probabilities occur more often than those with low probabilities.

It has been nearly a quarter century since a comprehensive Pentagon study of the future role of simulation for national-security analysis, planning, and operations, identified "agent-based models with emergent behaviors" as holding particular promise (Committee on Technology for Future Naval Forces-National Research Council, 1997). That study recognized how difficult it would be to wed high-quality social science to the rapid growth in computational capacities; how much would need to be invested; and how many funded projects would fail. It predicted that between 2010 and 2015 some agent-based, mission-domain models would be deployed and that by 2015 agent-based simulations would play a key role in system configurations for policy-planning purposes. Those forecasts were prescient. We are now within the zone of the future that those analysts imagined, even as the unfamiliarity, urgency, and complexity of the quadrant 2 challenges we face put an even higher premium on heeding their advice.

Conflict of Interest

The authors declare no conflict of interest.

ORCID

Ian S. Lustick https://orcid.org/0000-0001-7526-2679
Philip E. Tetlock https://orcid.org/0000-0002-3199-0292

ENDNOTES

[1] We offer this matrix as a heuristic device for classifying logically distinct types of problems encountered by analysts. Movement from one quadrant to another, or within one quadrant, reflects continuous variation in data availability or confidence in theoretical priors,

[2] On the actual, very strong, correlation between the importation of lemons into the United States from Mexico and American auto accident fatalities see Lowe (2009).

[3] The outstanding contemporary example of such a guru was Andrew Marshall, whose hunch early in the Cold War that Kremlin strategies were rational and that the US would profit from imitating them established his pronouncements and analytic techniques as authoritative for generations of US analysts. Marshall served as the director of the Defense Department's Office of Net Assessment for an astounding 42 years, from 1973 to 2015 (Evans, 2019; Krepinevich & Watts, 2015).

[4] Considerable attention has been paid to the use of table-top or role-playing simulations to address geopolitical questions in pedagogical or policy analysis contexts. The primary outlet for diverse applications of computational simulation techniques is the *Journal of Artificial Societies and Social Simulation*. As of October 2020, not one of the 250 articles in this journal reports research deploying computer simulation to address specific foreign-policy or national-security problems, although several articles do use simulation models to answer generic questions pertaining to geopolitics.

[5] For an example of using game theory to generate categories of scenario outcomes see Kydd and Straus (2013). For an intelligent though revealingly unsatisfying exploration of the difficulty of using such techniques to anticipate unusual, surprising, or rare events see King and Zeng (2001).

[6] ICEWS was one of the most successful of the many failures launched by different agencies within the intelligence community since early 2000's, each intended to develop social science-based tools, not only to forecast, but to help thwart, mitigate, or exploit threats and opportunities. These included Technologies for the Applications of Social Computing (TASC--renamed, in 2009, "Computational Social Science Experimental Proving Ground"); COMPOEX, PCAS, Pathways (CTTSO); Athena; MESA; SHARP; and the project reported on below, VSAFT. Some of these and many similar programs were developed under the umbrella of the Human Social Culture and Behavior Modeling Program in the Office of the Secretary of Defense from 2008–2013. See Mitre Corporation, Progress and Promise:Research and Engineering for Human Sociocultural Behavior Capability in the U.S. Department of Defense (June 2013) (Boiney & Foster, 2013; Egeth et al., 2014).

[7] ICEWS technologies for coding and tracking trends in event and sentiment data were transitioned.

[8] For detailed illustration of the federation of separate theories to test rival hypotheses about the causes of secessionism see Lustick et al. (2004).

[9] Systems dynamics, stock-and-flow computer simulation models can produce large numbers of randomly perturbed, unique trajectories, but have difficulty integrating propositions involving distinctive and changing interaction topologies and closely linked processes operating at different levels of analysis (O'Brien, 2010).

[10] For an excellent introduction to the concept and a useful theory of the conditions under which "emergence" emerges see Holland (1998).

[11] Each of the simulation exercises (and many more) were conducted with partial funding by different agencies within the US government. None of the work, and none of the information used, was classified. Much of it has been published and is cited. All were developed by Lustick and his team of researchers and modelers within the same overall research program. Development of the techniques and modules that were developed to produce VSAFT began in the mid-1990s, with funding from the Carnegie Corporation and the National Science Foundation, leading to preparation of an agent-based modeling interface, first called "ABIR" and then PS-I. Beginning with highly abstract simulations, work progressed toward generic models—renderings of domains of substantive theoretical and policy interest but not designed to correspond to specific times or places. Based on lessons learned and the effectiveness of generic models, more complex virtualization models were built using real-world data as inputs. VSAFT was the apparatus used to build the virtualizations, experiment with them, and analyze the results. For more on this research program and its results see Lustick et al. (2017); Lustick, 2002; Lustick et al. (2004); Lustick (2012).

[12] An account of the full range of VSAFT capabilities techniques applied to a difficult and detailed problem of real world concern is available in a report produced for the Holocaust Memorial Museum on the counterfactual problem of assessing what could have been done to reduce Syrian civil war atrocities. See Lustick et al. (2017).

[13] Detailed validation assessments of the previous months of model forecasts were regularly performed, including assignment of Brier scores. More general work on the challenge of "validation and verification" of social science models was also produced. See Alcorn et al. (2012) and Lustick and Tubin (2012). For data directly relevant to the Bangladesh study reported here, see Lustick Consulting, "VSAFT: Virtual Strategic Analysis & forecasting Tool: Quarterly Forecasting Verification and Validation Report" (June 2013) available on request from the authors. See also Lustick Consulting, "Badlands Extension Status Report 3," (May 10, 2016), available on request from the authors. See also Nicholson and Schmorrow (2012). The PS-I modeling platform along with operating instructions is freely downloadable at http://ps-i.sourceforge.net//. Models and protocols used for this and similar studies are available, along with data files. For detailed reporting on the Bangladesh model as representative of a large class of country models built with PS-I see Brandon Alcorn, Miguel Garces, and Ian S. Lustick, "Using Empirical Data to Test Theoretically Grounded Operationalizations of Protest in an Agent-Based Model," April 10, 2013, PS-I Modeling Repository; https://web.sas.upenn.edu/ilustick/2013/04/10/using-empirical-data-to-test-theoretically-grounded-operationalizations-of-protest-in-an-agent-based-model/.

[14] Instability measures are relative to the behavior of the Bangladesh model itself, not to absolute values for these categories applied across all country models. What is "calm" for Bangladesh might well be "intense politics" for Vietnam.

[15] This was what happened in the actual world. The Awami League government refused to dissolve in favor of a caretaker regime and instead supervised elections held on January 4, 2014.

[16] That is to say that the area of the red rectangle in the lower right-hand corner of Figure 2 is 29% larger than the red rectangle in the lower right-hand corner of Figure 3.

[17] In historical Bangladesh as well as in our model the Awami League has relied heavily on appeals to Bengali ethnonationalism. Both the Awami League and the more Islamically oriented BNP have traditionally maintained strong relations with the powerful business community in Bangladesh.

[18] On the validity of the model see sources listed in note 13.

[19] For detailed explication of both modules see Lustick et al. (2012).

[20] Syria Research Project, Simon Skjodt Center, Holocaust Memorial Museum (2016) https://www.ushmm.org/genocide-prevention/simon-skjodt-center/work/research/projects/syria-research/project.

[21] The intensity of the controversy caused the Holocaust Memorial Museum to remove the study from its website for a year. See Max

Fisher, "Holocaust Museum Tries Again on Contentious Syria Study," *New York Times* (December 19, 2017).

REFERENCES

Alcorn, B., Garces, M., & Lustick, I. S. (2012). Granular ABM simulations for operational use: Forecasting and what-if experiments with models of Kandahar and Kunduz. In D. M. Nicholson, & D. D. Schmorrow (Eds.), *Advances in design for cross-cultural activities part II* (pp. 24–33). CRC Press.

Aronoff, M. J. (1998). The politics of collective identity. *Reviews in Anthropology*, *27*(1), 71–85. https://doi.org/10.1080/00988157.1998.9978190

Betts, R. K. (2007). *Enemies of intelligence: Knowledge and power in American national security*. Columbia University Press.

Boiney, J., & Foster, D. (2013). *Progress and promise: Research and engineering for human sociocultural behavior capability in the US Department of Defense*. MITRE CORP MCLEAN VA.

Bradfield, R., Wright, G., Burt, G., Cairns, G., & Van Der Heijden, K. (2005). The origins and evolution of scenario techniques in long range business planning. *Futures*, *37*(8), 795–812. https://doi.org/10.1016/j.futures.2005.01.003

Brass, P. R., & Richard, P. (1980). Ethnic groups and nationalities: The formation, persistence, and transformation of ethnic identities over time. In P. F. Sugar (Ed.), *Ethnic diversity and conflict in Eastern Europe* (pp. 1–68). School of Oriental and African Studies, Centre of South Asian Studies.

Callaghy, T. M. (1984). *The state-society struggle: Zaire in comparative perspective*. Columbia University Press.

Chandra, K. (2012). *Constructivist theories of ethnic politics*. Oxford University Press.

Committee on Technology for Future Naval Forces-National Research Council. (1997). *Technology for the United States navy and marine corps, 2000–2035: Becoming a 21st-century force: volume 9: modeling and simulation*. : National Academies Press.

Lustick Consulting (2015). *Lustick consulting's ABM verification & validation*. Retrieved from https://cpb-us-w2.wpmucdn.com/web.sas.upenn.edu/dist/7/497/files/2016/06/LC_VandV_MiguelGarces.pdf

Coser, L. A. (1956). *The functions of social conflict*. Free Press.

Dahl, R. A. (1961). *Who governs?: Democracy and power in an American city*. Yale University Press.

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing theory through simulation methods. *Academy of Management Review*, *32*(2), 480–499. https://doi.org/10.5465/AMR.2007.24351453

Davis, P. K., Bankes, S. C., & Egner, M. (2007). *Enhancing strategic planning with massive scenario generation: Theory and experiments*. Rand Corporation.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Defense Modeling & Simulation Coordination Office (2020). VV&A. Retrieved from https://vva.msco.mil/

Dunning, T., & Harrison, L. (2010). Cross-cutting cleavages and ethnic voting: An experimental study of cousinage in Mali. *American Political Science Review*, *104*(1), 21–39. https://doi.org/10.1017/S0003055409990311

Easton, D. (1965). *A systems analysis of political life*. John Wiley & Sons.

Egeth, J., Klein, G. L., & Schmorrow, D. (2014). *Sociocultural behavior sensemaking: State of the art in understanding the operational environment*. MITRE CORP MCLEAN VA.

Evans, M. (2019). Reflections on an American seer: Andrew W. Marshall and the mind of the strategist. *Australian Journal of Defence and Strategic Studies*, *1*(1), 99–112.Retrieved from https://www.defence.gov.au/adc/Publications/AJDSS/documents/volume1-issue1/Comm3-Reflections-on-an-American-seer.pdf

Evans, P., Rueschemeyer, D., & Skocpol, T. (1985). *Bringing the state back in*. Cambridge University Press.

Ferguson, N. (1997). *Virtual history: Alternatives and counterfactuals*. Basic Books.

Fisher, M. (2017). Holocaust Museum Tries Again on Contentious Syria Study. New York Times.

Greenhill, B., Ward, M. D., & Sacks, A. (2011). The separation plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science*, *55*(4), 991–1002. https://doi.org/10.1111/j.1540-5907.2011.00525.x

Harrison, J. R., Lin, Z., Carroll, G. R., & Carley, K. M. (2007). Simulation Modeling in Organizational and Management Research. *Academy of Management Review*, *32*(4), 1229–1245. https://doi.org/10.5465/amr.2007.26586485

Hayden, M. V. (2017). *Playing to the edge: American intelligence in the age of terror*. Penguin.

Hayden, M. V. (2019). *The assault on intelligence: American national security in an age of lies*. Penguin Books.

Hitz, F. P., & Weiss, B. J. (2004). Helping the CIA and FBI connect the dots in the war on terror. *International Journal of Intelligence and Counter I*, *17*(1), 1–41. https://doi.org/10.1080/08850600490252641

Holland, J. (1998). *From chaos to order*. Addison-Wesley.

Huss, W. R., & Honton, E. J. (1987). Scenario planning-What style should you use? *Long Range Planning*, *20*(4), 21–29. https://doi.org/10.1016/0024-6301(87)90152-X

Jervis, R. (2010). *Why intelligence fails: Lessons from the Iranian Revolution and the Iraq War*. Cornell University Press.

Kahneman, D. (2011). *Thinking fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., Sunstein, C., & Sibony, O. (2021). *Noise*. Farrar, Strauss & Giroux.

Keil, F. C. (2005). Explanation and understanding. *Annual Review of Psychology*, *57*(1), 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Klein, A. (2003). The man who saw the future. Retrieved from https://www.strategy-business.com/article/8220?gko=4447f

Kowert, P., & Legro, J. (1996). Norms, identity, and their limits: A theoretical reprise. In P. J. Katzenstein (Ed.), *The culture of national security: Norms and identity in world politics* (pp. 451–497). Columbia University Press.

Krepinevich, A. F., & Watts, B. D. (2015). *The last warrior: Andrew Marshall and the shaping of modern American defense strategy*. Basic Books.

Kydd, A. H., & Straus, S. (2013). The road to hell? Third-party intervention to prevent atrocities. *American Journal of Political Science*, *57*(3), 673–684. https://doi.org/10.1111/ajps.12009

Laitin, D. D. (1998). *Identity in formation*. Cornell University Press.

Lipset, S. M. (1960). *Political man. The social bases of politics garden city*. Doubleday.

Lowe, D. (2009). Mexican lemons to the rescue. Retrieved from https://blogs.sciencemag.org/pipeline/archives/2009/04/01/mexican_lemons_to_the_rescue

Lustick, I. S. (1993). *Unsettled States, disputed lands: Britain and Ireland, France and Algeria, Israel and the West Bank-Gaza*. Cornell University Press.

Lustick, I. S. (1996). History, historiography, and political science: Multiple historical records and the problem of selection bias. *American Political Science Review*, *90*(3), 605–618. https://doi.org/10.2307/2082612

Lustick, I. S. (2002). PS-I: A user-friendly agent-based modeling platform for testing theories of political identity and political stability. *Journal of Artificial Societies and Social Simulation*, *5*(3).

Lustick, I. S. (2006). *Trapped in the war on terror*, Philadelphia, PA: University of Pennsylvania Press.

Lustick, I. S. (2012). Deploying constructivism for the analysis of rare events: How possible is the emergence of "Punjabistan". In K. Chandra (Ed.), *Constructivist Theories of Ethnic Politics* (pp. 422–451). Oxford University Press.

Lustick, I. S. (2016). Computer Assisted Agent-Based Modeling for Intelligence Purposes: The Actual, the Probable, the Plausible, and the Impossible. https://sites.nationalacademies.org/cs/groups/depssite/documents/webpage/deps_175779.pdf

Lustick, I. S., Alcorn, B., Garces, M., & Ruvinsky, A. (2012). From theory to simulation: The dynamic political hierarchy in country virtualisation models. *Journal of Experimental & Theoretical Artificial Intelligence*, *24*(3), 279–299. https://doi.org/10.1080/0952813X.2012.693841

Lustick, I. S., Garces, M., & McCauley, T. (2017). *An agent-based model of counterfactual opportunities for reducing atrocities in Syria, 2011–2014*. Simon-Skjodt Center for the Prevention of Genocide of the United States Holocaust Memorial Museum. https://cpb-us-w2.wpmucdn.com/web.sas.upenn.edu/dist/7/497/files/2019/04/AnAgentBasedModelofCounterfactualOpportunitiesforReducingAtrocitiesinSyria2011_2014-189mz89.pdf.

Lustick, I. S., & Miodownik, D. (2009). Abstractions, ensembles, and virtualizations: Simplicity and complexity in agent-based modeling. *Comparative Politics*, *41*(2), 223–244. https://doi.org/10.5129/001041509X12911362972070

Lustick, I. S., Miodownik, D., & Eidelson, R. J. (2004). Secessionism in multicultural states: Does sharing power prevent or encourage it? *American Political Science Review*, *98*(2), 209–229. https://doi.org/10.1017/S0003055404001108

Lustick, I. S., & Tubin, M. R. (2012). Verification as a form of validation: Deepening theory to broaden application of DOD protocols to the social sciences. In D. M. Nicholson, & D. D. Schmorrow (Eds.), *Advances in design for cross-cultural activities part II* (pp. 158–167). CRC Press.

Migdal, J. S. (2001). *State in society: Studying how states and societies transform and constitute one another*, Cambridge, UK: Cambridge University Press.

Nagel, J. (1994). Constructing ethnicity: Creating and recreating ethnic identity and culture. *Social Problems*, *41*(1), 152–176. https://doi.org/10.2307/3096847

Nettl, J. P. (1968). The state as a conceptual variable. *World Politics*, *20*, 559. https://doi.org/10.2307/2009684

North, D. C. (1981). *Structure and change in economic history*, New York: WW Norton.

O'Brien, S. P. (2010). Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, *12*(1), 87–104. https://doi.org/10.1111/j.1468-2486.2009.00914.x

Office of the Spokesperson (2014).Joint statement of the third U.S.-Bangladesh partnership dialogue. Retrieved from https://2009-2017.state.gov/r/pa/prs/ps/2014/10/233506.htm

Posner, D. N. (2005). *Institutions and ethnic politics in Africa*, New York: Cambridge University Press.

Ramírez, R., Österman, R., & Grönquist, D. (2013). Scenarios and early warnings as dynamic capabilities to frame managerial attention. *Technological Forecasting and Social Change*, *80*(4), 825–838. https://doi.org/10.1016/j.techfore.2012.10.029

Riaz, A. (2014). A crisis of democracy in Bangladesh. *Current History*, *112*(751), 150–156. https://doi.org/10.1525/curh.2014.113.762.150

Scheidel, W. (2019). *Escape from Rome: The failure of empire and the road to prosperity*, Princeton: Princeton University Press.

Schrodt, P. (2015). Seven observations on the newly released ICEWS data. Retrieved from https://asecondmouse.wordpress.com/2015/03/30/seven-observations-on-the-newly-released-icews-data/

Snowden, E. (2019). *Permanent record*, New York: Metropolitan Books.

Spaniol, M. J., & Rowland, N. J. (2019). Defining scenario. *Futures & Foresight Science*, *1*(1), e3–https://doi.org/10.1002/ffo2.3

Steinberger, M. (2020). The all-seeing eye. *The New York times Magazine*, *26–31*, 38–42.

Syria Research Project, Simon Skjodt Center, Holocaust Memorial Museum. (2016) https://www.ushmm.org/genocide-prevention/simon-skjodt-center/work/research/projects/syria-research/project

Tetlock, P. E. (2017). *Expert political judgment: How good is it? How can we know?*. Princeton University Press.

Tetlock, P. E., & Belkin, A. (1996). *Counterfactual thought experiments in world politics*. Princeton University Press.

Tetlock, P. E., Lebow, R. N., & Parker, G. (2006). *Unmaking the West: What-if scenarios that rewrite world history*. University of Michigan Press.

Tetlock, P. E., Lebow, R. N., & Parker, G. (2016). *Unmaking the west*. University of Michigan Press.

Tetlock, P. E., & Mellers, B. A. (2011). Intelligent management of intelligence agencies: Beyond accountability ping-pong. *American Psychologist*, *66*(6), 542. https://doi.org/10.1037/a0024285

Tilly, C. (1975). *The formation of national states in Western Europe*, Princeton: Princeton University Press.

Tsebelis, G. (1990). *Nested games: Rational choice in comparative politics* (Vol. *18*), Berkeley: University of California Press.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547. https://doi.org/10.1037/0033-295X.101.4.547

Vaughn, B. (2010). *Bangladesh: Political and strategic developments and US interests*. DIANE Publishing.

Wohlstetter, R. (1962). *Pearl Harbor: Warning and decision*. Stanford University Press.

Wright, G., Cairns, G., & Bradfield, R. (2013). Scenario methodology: New developments in theory and practice. Introduction to the Special Issue. *Technological Forecasting and Social Change*, *80*(4), 561–565. https://doi.org/10.1016/j.techfore.2012.11.011

How to cite this article: Lustick I. S., Tetlock P. E. The simulation manifesto: The limits of brute-force empiricism in geopolitical forecasting. *Futures & Foresight Science* 2021;e64. https://doi.org/10.1002/ffo2.64