A Conceptual Framework for AI-Augmented Statecraft

Project on Computational Statecraft Whitepaper

Charles Pozniak

AUGUST 2026

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

Abstract

This whitepaper lays the foundation for the Project on Computational Statecraft, an interdisciplinary research initiative at Harvard's Belfer Center dedicated to bringing computational methods to the practice of statecraft. Diplomatic practice cannot be reliably automated end to end by AI. Informed by diplomatic practice, computer science, and the social sciences, we argue that a disciplined, task-level approach to augmented statecraft can deliver material improvements to the speed, quality, range, and novelty of technical support available to human decisionmakers. The practices of statecraft, while irreducibly complex in aggregate, decompose into tasks with differing tractability and evaluability. We introduce a task-based architecture for identifying where AI systems can support human teams, and sketch which human judgments should remain decisive. This paper lays the foundation for methods, evaluations, and institutional designs required to safely and ethically expand the verifiable frontier.

Table of Contents

Acknowledgments

This paper is the first output of the Project on Computational Statecraft, housed within the Belfer Center's Program on Emerging Technology, Scientific Advancement, and Global Policy. It would not have been possible without the support of many people inside and outside Harvard University.

Meghan O'Sullivan, director of the Belfer Center, has supported this project from its earliest conception. Michael McQuade, the Program's director, and Rebekah Reed, our associate director, have been demanding and generous interlocutors and mentors throughout, while Kayleigh Lawson has supported much of the operational and logistical load. I am also grateful to the senior practitioners among the Center's fellows and faculty who agreed to be interviewed anonymously: their candor shaped the project's questions from the outset.

The project was conceived in partnership with two alumni of the Center, former Senior Fellow Carme Artigas and former Postdoctoral Fellow Anatoly Levshin. Each member of the team since has contributed to the ideas in this paper: Sameera Salari, Alisha Rajan, Jeba Sania, Jesse Berliner-Sachs, Moritz Hagemann, and Theo Lebryk. Two Technology and Geopolitics fellows, Josh Entsminger and Filippo Blancato, sharpened the arguments at several critical junctures.

Above all, this project would not exist without Slavina Ancheva, our partner in this work. Her unflappable temperament and joie de vivre carried the team through its early stages, and her intellectual input is present on every page.

For a project about the use of AI in statecraft, a word on our own use of it is owed. The team has used a range of open-source and proprietary models, including Claude and ChatGPT, to support a range of tasks, such as literature search and the organization of notes, as well as in the writing and presentation. Every claim, citation, fact, and figure has been checked against its source. The argument and the words are ours; the responsibility for them, and all errors, are mine

Introduction

This whitepaper lays the foundation for the Project on Computational Statecraft, an interdisciplinary research initiative at Harvard's Belfer Center dedicated to bringing computational methods to the practice of statecraft in an effective and ethical manner. Realizing the promise of AI in strategic settings requires interdisciplinary collaboration among diplomatic practitioners, social scientists, and technologists. The conceptual approach in this paper bridges the formal tools of computer science and statistics with the bodies of theory that already structure existing diplomatic practices.[1]

AI systems should not be framed as autonomous actors, but as a set of bounded tools whose usefulness depends on how well their work can be verified at the task level. Though game-playing AI successes have been striking, strategic geopolitical environments differ in several key dimensions that limit wholesale autonomous play. Though these challenges are limiting at the system level, they do not necessarily impact every subtask within the practices of statecraft. The responsible path forward, this whitepaper argues, is scoped augmentation as opposed to end-to-end automation.

Recent improvements in AI capabilities have opened a new space for technology companies providing decision support in high-stakes settings, but concrete applications are often ill-defined. Before 2020, a substantial strand of the academic literature focused on the foundations of automated negotiating agents, but the increased capabilities of LLMs have produced highly versatile systems that can represent and manipulate complex linguistic structures (Brown et al. 2020; Vaswani et al. 2017).

The practitioners of statecraft should be cautious of unverified claims about the utility of AI, specifically LLMs, in diplomatic settings. The potential benefits may be large, but the incentives for accurate public evaluation are currently misaligned: as leading AI labs have identified value in large government contracts, neither party wants to admit competitive (dis)advantages publicly.[2] For those offerings to add value to diplomacy, practitioners must have sufficient technical knowledge to understand the strengths and limitations of these new tools.

Strategic diplomatic environments differ from games because the rules, actions, and objectives are not fixed. These problems fit broadly under three categories of technical challenge:

- The *representation* problem asks whether systems can model the strategic environment. In diplomacy the game's 'board' is unstable — the moves, actors, and rules are themselves changing. Formally characterizing these components is required for computing moves and evaluating their correctness.
- The *inference* problem asks whether systems can learn what each actor wants and why things happen. Preferences, constraints, and beliefs are difficult to recover from behavior that is often designed to conceal them.
- The *evaluation* problem asks whether we can define success and judge what success looks like. Formalizing objectives requires preserving the normative and political dimensions that give diplomatic outcomes legitimacy.

To handle these limitations, this paper breaks down complex processes into different families of tasks: setup, research, analysis, strategy, execution, and monitoring. The objective of such support is a partnership between human and machine intelligence that helps organizations see further, move faster, and

1 For example, game theory, international relations, negotiation theory, and decision analysis, as well as computer science and product design.

2 This includes large deals between the government and technology companies like Scale AI and Anthropic.

reason more carefully about the future.

Human judgment, interpersonal connection, and democratic accountability will continue to be at the heart of diplomatic processes. As such, the limits of delegation to machines are tighter than in comparable private sector settings. The next generation of tools must keep central the human decisionmakers who can bear responsibility for outcomes, while forging genuine improvements that ensure safety and reliability alongside speed and efficiency.

If well-scoped AI augmentation delivers on even a fraction of its promise, it will change how states pursue their international objectives. The long-term goal of this research agenda is to ensure that gains build human accountability within healthier institutions. AI could improve the information available to each party, decrease the likelihood of misunderstanding, or help to level the playing field between large and small delegations. Reckless adoption may be rewarded by competitive pressures in private markets and world politics. Our work aims to limit that recklessness, ensuring the role of technology is bounded by discipline and responsibility.

This whitepaper is written for a range of audiences, including the practitioners who deploy AI tools, the policymakers who govern them, the technologists who build them, and the researchers who will study them and their impact. It proceeds as follows:

- Section I explains why end-to-end automation fails in diplomatic settings and outlines the contours of augmentation,
- Section II proposes a modular taxonomy of existing tasks in the practices of statecraft,
- Section III maps task categories to AI method classes, and,
- the concluding section charts open research questions.

I. The Limits of Automation in Foreign Policy

This section examines why applying AI to statecraft is harder than applying it to games like Go or chess. It identifies three challenges that efforts must confront: representation, inference, and evaluation — and addresses the risks of both inaction and premature automation. It concludes by outlining the argument for augmenting statecraft at the task-level, designing and evaluating tools for specific parts of the process, rather than monolithic systems that attempt to automate diplomacy end-to-end.

For this whitepaper, *statecraft* refers to the construction and execution of strategies and tactics through which political actors shape their external environment and influence other actors in pursuit of political ends (Kaplan 1952; Schelling 1960).[3] Strategies tie political objectives to instruments and theories of change, while tactics are the moves through which these are deployed (von Clausewitz 1832). In practice, statecraft operates through available *repertoires* of practices (Adler and Pouliot 2011; Goddard et al. 2019). Recent work has shown how statecraft is increasingly conducted as much through institutions and transnational networks as it is with direct interstate bargaining.[4] Diplomacy is then best understood as an institutionalized repertoire

3 Though the concept of statecraft may seem diffuse, it is typically studied through adjacent literatures on topics like coercion, bargaining, diplomacy, institutions, and grand strategy. The cleanest distinction is to see statecraft as broader than diplomacy, but narrower than power or bargaining.

4 We adopt the definition of diplomacy as a historically and culturally contingent bundle of practices (Pouliot and Cornut 2015) and international practices as "competent performances" (Adler and Pouliot 2011); Constantinou et al. (2021) argue for pluralizing diplomatic practice.

within statecraft, rather than its synonym.[5]

Statecraft is treated as multiple repertoires of practices rather than a single problem (Goddard et al. 2019). This makes computational augmentation most defensible when aimed at supporting a range of specific practices, as opposed to the whole process.

Large language models (LLMs) may construct coherent logics and strategies of statecraft, but they are also subject to the considerations outlined above — this means they cannot be verified, or 'trained', on correctness. These models are strong at synthesizing information and proposing linguistically coherent strategies, but proposals must be verified against some ground truth. If outputs are produced faster than institutions can verify them, then there is a risk from both processing bottlenecks and the integration of unchecked analysis into decisions.

A Field in Formation

AI is already entering the practices of statecraft. Ministries and foreign offices are moving quickly to deployment, with government agencies building tools: the U.S. State Department has deployed *StateChat*, and the U.S. Army has built *CamoGPT* (Satter et al. 2025; Ringquist 2025). A 2026 survey of 485 negotiation practitioners found AI tools embedded in preparation, analysis, and drafting; practitioners are already reporting concerns over confidentiality, over-reliance, and a loss of judgment (Bruderlein 2026).

Policy attention, by contrast, has focused on AI as an object of statecraft — chips, diffusion, and governance (Sullivan and Feldman 2026). Its arrival as an instrument of statecraft has proceeded piecemeal, largely without systematic evaluation, but research is converging on this space along three main tracks.

First, model behavior is evaluated in closed-ended foreign policy scenarios or compared against the choices of national security experts (Rivera et al. 2024; Lamparth et al. 2024; Payne 2026). These studies reveal the tendencies of language models placed in strategic decision environments, but do not theorize which tasks should or should not be undertaken by models.

A second branch of research extends the game-playing tradition, building agents that can participate in multi-party bargaining games and negotiation testbeds, most prominently the board game Diplomacy (Meta Fundamental AI Research Diplomacy Team (FAIR)† et al. 2022)**.** Relatedly, AI-mediated deliberation has been found to produce group statements that participants prefer to those of human mediators (Tessler et al. 2024). A final branch studies institutions: digital diplomacy scholarship has started mapping AI-driven practice and practice-oriented work, particularly in peace mediation and negotiation (Manor 2026; Hirblinger 2022).

The older empirical tradition of geopolitical forecasting shares structural components with this whitepaper's proposal. Forecasting tournaments, such as those conducted under IARPA's *Aggregative Contingent Estimation* program, scored probabilistic estimates of political events against explicit resolution criteria and demonstrated that training, teaming, and aggregation could improve accuracy, producing measurable gains over unaided expert judgment (Mellers et al. 2014; Tetlock and Gardner 2015). This work supports a central premise of this whitepaper: that subtasks of strategic judgment can be effectively isolated

5 In the international system, power is allocated and contested through processes of negotiation and diplomacy (Lukes 2021; Barnett and Finnemore 2004). A rich literature examines these processes, particularly the origins of institutions and their functions in the modern order (Waltz 2010; Ikenberry 2009; D. A. Baldwin 1980).

and evaluated against external criteria, supporting improved team performance.

Despite this innovative work, the field lacks a connective layer linking task-level activities to computational methods and to standards of evaluation. Benchmarks have relied on stylized, high-level scenarios that lack the complexity of diplomatic work, while agents have mastered games whose mastery does not transfer to statecraft — as this whitepaper will explore (Jensen et al. 2025). The field is rich in promising findings, but those findings lack placement within institutional practice and shared criteria of evaluation.

This paper outlines that connective layer, grounding technical research in real-world practice. The taxonomy is designed to map decomposed tasks onto method classes and evaluation criteria, providing a common architecture in which technical tools can be placed within the repertoires of statecraft. As the capabilities of AI systems continue to evolve, the practice and study of world politics need methods to establish the value of each new tool.

The Boundary of Formalization

Game theory and bargaining theory provide the most mature formal tools for strategic analysis in institutional settings. Foundational work on bargaining, deterrence, incomplete information, and two-level games has shaped the study of diplomatic strategy and conflict (von Neumann and Morgenstern 1944; Nash 1950; Harsanyi 1967). Decision analysis complements these traditions by focusing on how actors should make choices under uncertainty, using subjective probabilities, expected utilities, and explicit value tradeoffs, allowing actors to generate practical guidance amidst complexity (Lax and Sebenius 1986; Keeney and Raiffa 1993; Raiffa et al. 2002). In contrast to approaches oriented mainly toward equilibrium characterization or descriptive explanation, decision analysis is explicitly prescriptive. It breaks complex strategic problems into tractable components, including objectives, uncertainties, payoffs, interdependence, repeated interaction, and institutional design under specified constraints.

The Automated Negotiating Agents Competition has advanced research on bilateral multi-issue negotiation since 2010, producing increasingly sophisticated agents in structured settings (Baarslag et al. 2015). Work on the *Human-Agent League* and related platforms has clarified the gap between agent-agent benchmarks and negotiation with human counterparts — a gap that has narrowed but still remains informative (Mell et al. 2018).[6]

Recent AI achievements illuminate where the boundary of reliable formalization lies. Meta's CICERO system achieved human-level performance in the game Diplomacy[7] by combining a dialogue model with strategic planning and reinforcement learning (RL), providing us with evidence that language-conditioned strategic agents can operate in structured negotiation games (Meta Fundamental AI Research Diplomacy Team (FAIR)† et al. 2022).

While CICERO was a significant achievement, the gap between board games and statecraft is instructive. The game Diplomacy has fixed rules, known players, a discrete action space, and a well-defined objective

6 Before the latest increases in AI capabilities, humans would exploit the agents' reliance on well-defined utility functions, introducing ambiguity, emotion, and cultural dimensions that the agents struggle to reliably evaluate (Aydoğan et al. 2021; Lin et al. 2014; Mell and Gratch 2017).

7 Diplomacy is a board game set in pre-World War I Europe that tests strategy and negotiation. Players control great powers and seek territory through negotiation, alliances, deception, and coordinated military action. The agreements are non-binding and moves are resolved simultaneously, making it a canonical game for studying strategy, communication, and bargaining.

function; CICERO did not have to grapple with the central problems of formalizing actual negotiations (representation, inference, and evaluation).[8] Real diplomatic settings superficially resemble well-structured games, but their complexity lies in the features that board games hold constant.

As LLMs produce increasingly sophisticated analyses, the central question becomes verification: systematically interrogating whether a given contribution supports an actor's goals. In closed domains, the correctness of broad strategic recommendations can be operationalized *ex ante* through externally fixed objectives and rules and computed positional advantage. For example, a chess engine's recommended move can be evaluated using search and self-play, and, in some endgames, against ground truth. These evaluations are imperfect but stable because the rules and the objectives are fixed, allowing players to see whether a certain move brings the player to a more favorable position. Verification is possible because the evaluative criteria are both stable and external to the system.

Holistic, end-to-end diplomatic outputs ("automated" negotiators) currently resist definitive *ex-ante* verification for the reasons this section will explore. Strategic assessments cannot be checked against ground truth, preferences are strategically concealed, and the relevant counterfactuals are unobservable. The evaluative criteria for a "correct" recommendation are themselves politically contested and strategically manipulable.

Despite the limits of holistic strategic recommendations, many constituent sub-components can be verified. Components such as factual premises, logical coherence, coverage, and consistency can be validated before action is taken. This task-based decomposition can recover verifiability for individual components of a process whose aggregate output remains unverifiable in advance.

Task-based decomposition

The logic of decomposition has roots in systems theory: Complex systems are amenable to analysis precisely to the extent that they are decomposable into semi-independent modules (Simon 1962). Decomposing the processes of statecraft into its constituent tasks allows for verification (*ex-ante* and *ex-post*) at the subcomponent level. A taxonomic decomposition can recover partial verifiability by isolating the subtasks that have stable and external evaluation criteria. For example:

1. A system that retrieves and synthesizes background research can be evaluated for accuracy or coverage against a documented range of sources;
2. A system that generates negotiation options can be reviewed for range and plausibility by domain experts — a weaker check than verification;
3. A system that monitors compliance can be scored against specified, observable indicators.

In each of these examples, tractable evaluations are only possible because decomposition has sufficiently separated each task from the contested core or interconnected assumptions. If an AI system tries to take multiple steps at once, then assumptions over contested values, ground truth, or forecasting models can be carried through and compound over time. Each of these ambiguities requires accountable human judgment to resolve. The holistic 'correctness' of a broad end-to-end strategy cannot be evaluated

8 Other simplifications include a fixed player set, common knowledge of fixed rules, a single well-defined objective, a discrete action space, and no implementation uncertainty.

like this.

Well-evaluated components can have errors that interact and propagate when recompiled at the system-level, particularly in the absence of human interventions. The main risk here is an *interface* error: a subtask may be locally well-evaluated while still producing outputs that distort a broader strategic process. In systems theory, decomposition is faithful when the interactions between subtasks are explicitly specified, allowing for each locally valid output to compose without distortion (C. Y. Baldwin and Clark 2000). The relationships between subtasks should be preserved during recombination, whether each input is pooled, sequentially continued, or reciprocally interdependent (Thompson 1967). The problem of faithful recombination is a central open challenge of the approach that we outline. As there is no guarantee that each verifiable sub-task will aggregate up to a verifiable system, it is essential that we have human teams both executing on tasks beyond machine capability and recombining each of the decomposed subtasks. Human oversight can provide risk mitigation, but no guarantees to correctness. Any designed system, however, should ensure consistent interdependencies across the (loosely coupled) network (Brusoni et al. 2001).

As models improve, AI systems will likely be able to competently handle a broader chunk of relevant subtasks, narrowing the remaining areas of strategic judgment. As such, institutions at the heart of statecraft should ensure that their systems can adequately preserve human judgment, visibility into the system, and mechanisms of democratic accountability amidst the increasing capability. Participants in world politics will need to maintain accountability over the irreducibly strategic, while building trust and institutional capacity incrementally.

Not all mechanisms for model evaluation are equal, with several distinct tiers of evaluability. This includes:

(i) externally checkable outputs, or indicators (this captures most research or monitoring tasks, with some scope for assessing the correctness of analysis),
(ii) expert-auditable outputs, where reviewers can assess validity but without a ground truth to compare (this includes most analysis tasks or option generation processes),
(iii) non-verifiable judgment calls involving contested value or unknowable predictions (this captures most strategic judgment calls, or live execution tasks).

Evaluability alone does not defend the decision to delegate a task. A system may produce accurate and auditable outputs while still making some judgment over contested values or objectives. Deployment decisions should balance both the evaluability of the output and the institutional setting within which the decision will be made.

Three Challenges

This section examines the resulting three challenges from the computer science and computational social science literature:

1. The *representation* problem asks whether we can model the strategic environment with enough faithfulness to the underlying complexities of reality: whether a model's states, transitions, and action sets can be formally captured in ways that support computation.
2. The *inference* problem asks whether we can find out what we need to know: whether preferences, constraints, and beliefs can be uncovered from observable behavior.
3. The *evaluation* problem asks whether we can define success: whether objectives can be formalized

without sacrificing the contested normative and political dimensions at the heart of diplomatic practices.

The Representation Problem

To meaningfully apply computational methods to diplomatic settings, we must be able to represent strategic environments in formal terms — that is, as variables, relationships and rules that AI models and systems can process. This requires specifying *states* (who the actors are, what they believe, what constraints they face), *actions* (what moves are available), and *transitions* (how actions change states, which states are likely to follow others).[9]

Institutional endogeneity

The lack of consistent constraints in world politics means that game-playing computational approaches are often inapplicable; these methods treat institutional structures as fixed, like the rules of a game, as opposed to dynamic and non-binding. Institutions structure political and economic interaction by lowering transaction costs, reducing uncertainty, and constraining feasible actions (Coase 1937; North 1991; Keohane 1988). The rational design literature explains why states create particular rules as strategic responses to cooperation problems (Koremenos et al. 2001). This scholarship also demonstrates that institutional design is itself strategic as actors anticipate how rules will constrain future bargaining and design accordingly.

The rules governing diplomatic interaction are endogenous variables, with negotiations creating, modifying, and dissolving the institutions within which they occur. For example, the 2015 Paris Agreement's ratchet mechanism for progressively tightening climate commitments was itself a negotiated institutional innovation that changed the structure of future climate diplomacy. The evolution of WTO dispute settlement procedures, the creation of ad hoc coalitions outside formal treaty structures, and the erosion of arms control regimes all illustrate the same dynamic: Institutional rules are part of what is being contested and (re)constructed.

The WTO Appellate Body crisis illustrates how international institutions can have their own rules endogenously transformed by participating states. Since 2019, the WTO Appellate Body has been unable to properly function due to losing its three-member quorum.[10] Although there are formal rules for the system, the scope of those rules has evolved through strategic non-cooperation. This is a move that sits outside of the institution's founding "rules" but has transformed the space of potential subsequent moves. An AI system that searches over allowable institutional moves at its initiation, like in a board game, would not have discovered this.

Action space complexity

Unlike games with externally specified legal moves, statecraft has no bounded action set. Systems that

9 Markov Decision Processes (MDPs) provide a canonical framework for sequential decision-making under uncertainty; Partially Observed MDPs (POMDPs) extend this in settings of hidden information, while stochastic games or partially observable stochastic games extend it to multi-agent settings (Bellman 1957; Sutton and Barto 2018; Littman 1994). Bayesian games model private information and strategic interaction, while structural causal models and agent-based models offer complementary tools for causal representation and simulation when equilibrium or Markov assumptions are inappropriate.

10 Dispute settlement has not stopped entirely: panels continue to issue reports, and members have improvised workarounds such as Article 25 arbitration and the Multi-Party Interim Appeal Arbitration Arrangement (MPIA).

combine reinforcement learning, search, and language modules have reached human-level performance in complex games such as chess and Go, as well as video games like Atari and StarCraft II (Mnih et al. 2015; Silver et al. 2018; Vinyals et al. 2019). But the legal moves in these games are discrete and externally specified.

In diplomatic settings, a negotiator can propose novel treaty language, invoke new precedents, or take actions outside the formal negotiation. The action space is not usefully enumerable ex ante. Institutional context, political feasibility, and cognitive constraints impose practical bounds that formal models can only imperfectly approximate.

This complexity of the action space has formal consequences. Pruning the action space to a manageable set of appropriate moves is possible, but it must be informed by domain-specific institutional knowledge rather than derived from game-tree statistics alone. Computational systems do not have to enumerate all possible actions. Instead, they should be able to generate a diverse subset of strategies that includes unconventional options an experienced team might overlook.

In world politics, unconventional moves sometimes succeed because they violate expectations (Jervis 1976; Handel 1981). AlphaGo's celebrated move 37 was valuable because it lay outside the space of moves humans would consider but still within the space of possible moves (Silver et al. 2016). Any such pruning must distinguish moves that are genuinely unpermitted from those that are surprising but feasible.

Multiparty complexity

Much of the canonical game-theoretic work on crisis bargaining begins with two-actor models, with multilateral negotiations making the strategic structure more fluid (Schelling 1960; Fearon 1995; Zartman and Berman 1982; Odell 2000). Complexity comes from the contested and evolving strategic structure, or "non-stationarity", as well as the increased number of players. Changing issue-based alignments are often contested and evolving; computational techniques, like multi-agent reinforcement learning, handle such fluidity poorly (Littman 1994; Shoham et al. 2007; Busoniu et al. 2008; Hernandez-Leal et al. 2019; Chalkiadakis et al. 2022).

Limited stationarity and path dependence

Even when institutional rules and actors are fixed, negotiation arrangements may differ due to the intervening path dependencies. Relationships accumulate history over time, with reputations and past concessions constraining future credibility (Axelrod 1984). Repeated game models capture how the shadow of the future shapes present choices, and reputation models formalize how past behavior constrains beliefs about future conduct. But these models typically assume stable preferences and fixed game structures. Many standard reinforcement learning and off-policy evaluation methods assume that transition and reward distributions are stable enough for past data to inform future decisions. When preferences, institutions, relationships, and issues evolve together, diplomatic settings challenge these stationarity assumptions.[11]

Any formal model of a strategic diplomatic environment will either respect these properties and pay the resulting complexity costs or violate them and pay in fidelity. There is no modeling strategy that avoids both costs simultaneously. The practical implication is that formalization is possible but inherently

11 The degree of non-stationarity varies, however: Some institutional features (voting rules, membership criteria, procedural norms) change slowly enough to serve as stable scaffolding for computational models, even as other parameters shift around them.

partial. Formal models will cover some subset of the state space with high fidelity and degrade outside that subset. We will not build a complete model for world politics. We can, however, characterize the boundary of each tool's validity with sufficient precision for practitioners. This is a problem that asks for calibrated incompleteness: models that know what they don't know.

This representation problem is a domain-specific instance of the general knowledge representation problem: how to represent a complex, changing world in a form that supports automated reasoning (Davis et al. 1993). The specific difficulties of representation in international relations correspond to known hard problems in computer science.[12] Each of these has a substantial technical literature that this project can draw upon and extend.

The Inference Problem

Even if it were possible to adequately model the strategic environment, we would face a second challenge: inferring the parameters — the key factors — that shape diplomatic behavior and outcomes. To provide useful decision support, a system should represent and update key beliefs, for example, about what actors want, the constraints they face, or what they believe about each other (Harsanyi 1967; Lake and Powell 1999). In principle, these can be inferred from observed behavior.[13] In practice, diplomatic behavior is often designed to both reveal and conceal, communicating credibility and resolve while maintaining bargaining advantage and strategic ambiguity.

Strategic misrepresentation

The foundational puzzle of rationalist theories of conflict is that war is costly and inefficient but still occurs frequently (Fearon 1995).[14] Explanations for this puzzle include private information, incentives to misrepresent, commitment problems, and issue indivisibility, with later work incorporating time horizons (Toft 2006). Importantly, the existence of a nonempty bargaining range is not automatic; if the stakes are perceived as existential or values genuinely incompatible, then mutually acceptable settlements may not exist at all (Fearon 1995). For this paper, the key implication is that the behaviors an AI system must interpret in the conduct of statecraft are often strategically constructed to obscure the information that inference would require.

Preferences can be inferred computationally from observed behavior, but current algorithms are narrow in statecraft-relevant applications (Ng and Russell 2000). This is done through techniques like inverse reinforcement learning (IRL), which depends on strong assumptions about the environment, actors, and their rewards. These techniques assume that observed actions reveal underlying reward functions. In canonical IRL problems, the goal is to recover a reward function under which observed behavior is optimal or near optimal. This is often underspecified as the same behavior could be optimal for many objectives. In the

12 Institutional endogeneity maps to the *frame problem* (what changes and what persists when an action is taken); unbounded action spaces map to the *open-world assumption* (the set of possible propositions is not fixed in advance); multi-party complexity maps to *multi-agent epistemic reasoning* (different agents may have incompatible but individually coherent world models); and limited stationarity maps to the problem of *belief revision*.

13 This form of massive inverse reinforcement learning happens in large, consumer-facing applications like Google Maps (Barnes et al. 2024).

14 If issues are divisible or compensable and actors can credibly commit to a settlement, then a mutually preferred bargaining range should exist under complete information.

conduct of statecraft, particularly during diplomatic bargaining, this identification problem is compounded because the observed behavior is itself strategic (Kydd 2005; Sartori 2005; Jervis 1976). Because actions are selected to influence future outcomes, observed behavior is mediated by incentives and constraints, rather than directly revealing beliefs or preferences. Understanding these complications is central for building effective systems that support the practices of statecraft.

The dynamics of statecraft can be more accurately characterized through multi-agent IRL or inverse game theory under asymmetric information. In cooperative settings, the informed actor may have incentives to teach or reveal the relevant reward function, allowing for the function to be inferred (Hadfield-Menell et al. 2016). Diplomatic bargaining often has the opposite structure: actors may have incentives to mask or manipulate beliefs through ambiguity. Non-cooperative settings are more difficult, but reward functions can sometimes be partially recovered in stylized cases, even when one agent obscures its objective from another agent (Zhang et al. 2019).

During the Iran nuclear negotiations that culminated in the 2015 JCPOA, the parties had to infer genuine red lines from stated positions. Outside observers and counterparties had to distinguish genuine constraints from negotiable positions. Iran's insistence on enrichment capacity could be a constraint reflecting domestic politics and technical path-dependence, or a bargaining position designed to extract concessions elsewhere (Samore 2015). Computational systems cannot resolve this ambiguity by collecting more passive data because the ambiguity is deliberately generated by the actors themselves (Jervis 1976). The difficulty in deciphering intent arises because actors have incentives to distort signals.

As AI capabilities increase, new mechanisms will be required that adapt to new technical capabilities to incentivize cooperation. This shifts the problem from statistical inference to mechanism and institutional design. Instead of attempting to recover information from more data alone, parties need arrangements under which informative signals are costly to fake or are beneficial to reveal.[15]

On defined forecasting benchmarks, data-driven models can outperform physics-based numerical models (Lam et al. 2023).[16] Commitment credibility, however, faces obstacles that weather does not: Actors strategically manipulate the signals a prediction system would need to learn from. Additionally, the deployment of such systems would alter signaling behavior, and the relevant cases to learn from number in the hundreds rather than billions.

Endogenous win-sets

Putnam's two-level games framework is foundational to diplomatic analysis, capturing how international negotiations are constrained by domestic ratification requirements (Putnam 1988). What Putnam characterizes as the "win-set" is the set of international agreements that would be domestically ratifiable and, therefore, determines what deals are feasible. In practice, leaders also try to shape the sets through narratives, meaning that win-sets are themselves endogenous, strategic instruments that leaders expand or contract based on bargaining incentives.

Leaders may invoke domestic constraints to extract concessions at the negotiating table, or manufacture

15 Commitments are made credible by factors like reputation, domestic political costs of reneging, relationship-specific investments, third-party guarantees, and the shadow of future interactions (Schelling 1960; Fearon 1995; Koremenos 2001).

16 DeepMind's GraphCast, trained on decades of reanalysis data, surpassed ECMWF's leading deterministic forecast system on most evaluated medium-range weather targets.

flexibility to enable deals that serve their interests. For example, the U.S. movement in and out of the Paris Climate Agreement illustrates the role of domestic political constraints. The 2017 announcement, 2020 withdrawal, 2021 reentry, and 2025-26 withdrawal each shifted beliefs about the durability of American climate commitments across administrations. Any estimation procedure must contend with the fact that observed boundaries of win sets are also strategic signals.

Data constraints

The underlying data for effective inference are uneven: rich in some domains (trade negotiations or UN voting) and thin in others (crisis diplomacy or backchannel communications). High-stakes negotiations occur rarely, with few cases comparable to the Cuban Missile Crisis, Camp David, or the Iran nuclear talks. The available data is also selection biased. Observers see only negotiations that reached public stages and have little evidence from failed negotiations or exploratory contacts.

In our preliminary interviews conducted for this project, senior practitioners suggested that negotiation success often depends on personal relationships that operate beneath the visible surface of state interests (Ancheva, forthcoming). This whitepaper focuses on process-level augmentation that underpins both the substantive analysis and the relational work — though the latter remains human-executed. The interpersonal dimension is the highest-risk area for automation and is where personal accountability and relationship management will remain critical.

Practitioners also have examples from their own experience that never entered the documentary record. This knowledge cannot be readily retrieved because it was never institutionally stored (Ancheva, forthcoming). The unrecorded judgments of diplomats include a range of subtle interpersonal and inferential skills. These skills might include recognizing a bluff, or an intuition on a shift in political momentum. Interpersonal expertise is consequential but invisible to systems that learn from documents. The experience of each individual diplomat — every win or rebuke over the course of a career— is a far richer learning signal than the outcome-level data or transcripts that we might be able to access after the fact.

The Evaluation Problem

Even if one solves the representation and inference problems, defining and evaluating success across multiple participants remains difficult. Classic optimization techniques need some objective that can be represented mathematically, such as profit or time, to then maximize or minimize. Modern optimization methods are sophisticated, but diplomatic outcomes resist evaluation in these terms for reasons that are normative and political, as well as technical.

Defining Success

A government may have many overlapping, often competing, objectives in a negotiation. Security, economic welfare, international status, domestic political survival, and normative legitimacy may all feature simultaneously. These may be weighted differently by different actors within the same government, and the weights are constantly shifting as circumstances evolve (Allison and Zelikow 1999).

Allison and Zelikow's analysis of the Cuban Missile Crisis demonstrated that even a single government's preferences reflect multiple lenses. The "rational actor" model treats the government as a unitary optimizing

actor; the "organizational process" model captures how routines shape options; the "bureaucratic politics" model captures how competing internal interests produce policy (Allison and Zelikow 1999). Because no single interest or objective is complete, synthesizing them into a comprehensible strategy remains a matter of qualitative judgment. There is no single objective function for an AI system to optimize as a holistic strategy, because there is no single actor, but locally specified metrics for scoped tasks remain possible.

Quantitative tools can serve many valuable functions in this process but cannot determine which contested political values should dominate when facing conflicting objectives. In this setting, tools may clarify tradeoffs or expose inconsistencies, but there are other constraints that prevent them from providing concrete answers. These include normative constraints, issues of legitimacy, and multiple audiences with strong —often conflicting — preferences.

Normative constraints and legitimacy

In a negotiation, some options are off-limits for reasons beyond strategic costs, particularly in peace agreements and legalization (Chayes and Chayes 1995; Fortna 2004; Abbott et al. 2000)**.** Violations of sovereignty, human rights, and international law may be strategically advantageous and still be impermissible for a given actor. Across the table, one side's moral sin might be another's calculated trade-off. Within a single team, one person's deepest value might be another's bargaining chip. Legitimacy is normatively constructed and concerns both outcomes and their justifications. Any AI system will need to make these distinctions explicit as opposed to aggregating all normative concerns into a single number or bucket.

When facing normative trade-offs, the role of an AI system may be like Isaiah Berlin's role of the moral philosopher: outlining costs and tradeoffs in each situation as opposed to giving a judgment. Berlin held that the moral philosopher's role is to lay out the values, issues, and forms of life in collision with one another, as opposed to adjudicating between them (Magee and Berlin 1978). This left both the choice and responsibility to each individual. When facing contested values, this should be seen as a best-case scenario for AI integration — though better models may help make the facts clearer, or consequences sharper, the choice and the responsibility are non-delegable.

Multiple audiences

Diplomatic communications address multiple audiences, publicly and privately. A message that signals resolve to one audience may signal stubbornness to another; a concession that builds trust with a counterpart may provoke domestic backlash. With multiple audiences, diplomats handle this complexity with real-time judgment under uncertainty (Putnam 1988). Reducing it to an objective function would require artificially fixing contestable political choices.

These problems with evaluation must be seriously considered with any augmentation design. AI tools may not be able to fully specify contested values but must make the structure of the problem transparent. This involves identifying the tradeoff frontiers between competing objectives and flagging cases where a proposed strategy performs well on one metric only by performing poorly on another. This provides decision support in the literal sense: supporting the human act of deciding what to value by clarifying what each choice of values entails.

Potential Benefits and Costs

Scoped computational tools can deliver genuine value to augmenting the existing practices of statecraft; we posit four dimensions to assess this value — speed, quality, range, and novelty. Achieving and validating these benefits will require systems designed with sufficient modularity and transparency to enable comparisons to benchmarks of existing activities.

- *Speed.* Research that currently takes weeks could be accomplished in hours with existing language model capabilities.[17] Faster preparation increases capacity for analysis and strategy, conferring early advantages to parties that adopt appropriate tools (Dell'Acqua et al. 2023; Noy and Zhang 2023).
- *Quality*. Broader source coverage and systematic analyses could reduce the gaps and inconsistencies that impact preparation under time constraints. Realizing quality gains will require disciplined institutional processes, but early evidence suggests that human-AI collaboration can increase creative problem-solving (Boussioux et al. 2024).
- *Range*. Computational tools can evaluate more scenarios and surface more possibilities than human teams alone working under time pressure. This expansion of known options is where augmentation may add significant value.
- *Novelty*. AI systems may expand the option space by exploring precedent from prior negotiations, constraints faced, and future paths in ways underexplored by human teams. The possibility of "move 37" moments, where unconventional options are found outside of human experience, is real, albeit currently rare. We see decision-relevant novelty as the longest-horizon payoff of this research agenda. Although current AI methods may present unconventional options, systems cannot reliably certify specific options as better or worse when it would be most consequential. Identifying the value of such novelty depends on strong evaluation mechanisms and strategic judgment under immense uncertainty.[18]

These dimensions of improvement should be scoped and measured against baseline performance of real-world human teams. Without benchmarking, factors like speed or accuracy cannot be accurately assessed, and claims of AI-augmented improvement may remain unfalsifiable. Baselines provide value to institutions, potentially revealing inefficiencies and inconsistencies that can be addressed independently of AI. Moreover, baseline data enables comparative evaluation across tools and institutions, allowing the field to distinguish genuine future advances from exaggerated vendor claims or confirmation bias.

The risks of premature deployment

The competitive dynamics of contemporary geopolitics create pressure toward deployment of AI tools. If adversaries or counterparts are automating diplomatic processes, the instinct to match their capabilities becomes difficult to resist (Horowitz 2018; Allen and Chan 2017). But deploying systems that appear competent while lacking substantive reliability creates specific harms, some of which are predictable (Pozniak and Sania 2026). The cause for caution is found throughout the adjacent empirical evidence. For example, off-the-shelf LLMs display erratic escalation patterns when making strategic decisions in simulated wargames, even in scenarios without conflict seeded (Rivera et al. 2024).

17 In adjacent knowledge-work settings, generative AI has brought material time savings on bounded writing and consulting tasks. Whether or not comparable gains are seen in diplomatic preparation should be treated as an open empirical question, benchmarked against existing research workflows.

18 Historical examples of institutional novelty include the first armed peacekeeping force, deployed as a diplomatic technique by the United Nations in the 1956 Suez Crisis, and the introduction of multilateral financial lending by the League of Nations as part of the Geneva reconstruction program for Austria (1922-23).

Mistakes in diplomatic contexts are costly. Errors, like misunderstandings or missteps, carry consequences with tangible human costs that can compound over time. Unlike consumer applications where errors are frequent and low-cost, diplomatic errors can permanently damage relationships or escalate conflicts.

The verification asymmetry expounded above compounds these risks. Current systems are trained to perform the surface features of competent analysis with linguistic fluency, presenting clear structure and appropriate citations. Outputs, therefore, may pass superficial review while potentially misjudging strategic dynamics in ways only domain expertise would detect.

Models are becoming incredibly capable at a range of qualitative knowledge tasks, such as complex reasoning, formal writing, and searching over large databases, though limitations still remain (Wei et al. 2023; Dell'Acqua et al. 2023; Liu et al. 2023)**.** These capabilities are, however, uneven and failure modes can be difficult to predict. A model's jagged competence profile is potentially dangerous if insufficiently understood and indiscriminately deployed. Systems that perform well on writing and summarization tasks may fail on adjacent tasks. Interpreting strategic signals and drafting commitments look mechanically similar to the summarization and formal writing that models do well, but they are far harder to generate, and to verify. Excess faith in a model's output due to adjacent competencies could introduce blind spots and risk.

The disempowerment of human teams is an important long-term risk with early evidence from a range of literatures, but no data to date to support a strong conclusion. Across industries from healthcare to law, evidence on the benefits and risks of decision delegation and automation is conflicting (Bainbridge 1983; Parasuraman and Manzey 2010). There is a risk that practitioners may become dependent on AI-generated analyses. Without rigorous analytical work, individuals may lose the capacity to evaluate outputs, and eventually to perform the analysis themselves, as the judgment that made delegation safe erodes through disuse (Endsley and Kiris 1995). Organizations that automate extensively may find, when systems fail, that human expertise has been eroded. Such organizations may find they then lack the institutional ability to detect and correct errors.

When multiple parties rely on similar AI systems, sufficiently correlated errors can induce systemic risk. Adjacent domains have explored how similar AI systems can produce simultaneous misreadings, coordinated escalations, or collective blindness to possibilities outside the training distribution (Daníelsson et al. 2022). The diversity of human decisions, for all its inefficiencies, provides a hedge against correlated failures that homogeneous AI systems may not be able to replicate (Hong and Page 2004).

Adversarial pressure on the technology stack.

The analysis of this whitepaper treats counterpart strategic misrepresentation primarily as an inference problem, but it is important to note that AI deployments can function as an additional surface for adversarial counterparts to attack. AI systems may be deliberately targeted through malicious cyberattacks. With decision-support systems, these attacks might attempt to subvert agents themselves, potentially through 'poisoning' training data or open-source information to steer agents off-course. These approaches are well-established the adversarial ML literature (Vassilev et al. 2025). In this domain, however, undetected attacks could be devastating — akin to having a 'double agent' implanted as a key advisor. As such, any AI risk assessment should treat cybersecurity deployment as a primary consideration.

The risks of inaction

Despite these concerns, the risks of inaction are material. Parties that use AI tools to improve their analytical capabilities, expand their strategic range, and accelerate their response times may hold advantages over those that do not. This could confer benefits across a range of dimensions in adversarial settings.

Beyond competitive pressure, existing bureaucratic processes in statecraft are often cumbersome, slow, and prone to their own systematic biases. Groupthink, availability heuristics, anchoring on familiar precedents, and the clustering of analyses around salient examples are documented challenges for human teams (Janis 1972; Tversky and Kahneman 1973; Welch 2000; Clement and Tse 2005). Choosing not to implement appropriate AI tools means accepting the limitations of current practice while forgoing potential improvements provided by advanced computational tools.

There is an opportunity cost if AI tools are not actively studied for applications in diplomatic settings. If scholars and institutions abstain from developing AI augmentation, the field will be defined by actors with less concern for rigor, safety, or the preservation of democratic accountability. Responsible development requires engagement, not withdrawal. Internal pressures to modernize from an ill-considered perspective may result in premature adoption with institutional downsides.

These system-level constraints motivate the task taxonomy developed in Section II.

II. The Task Taxonomy

Though there is a multitude of AI tools and systems that could potentially assist in the practices of statecraft, this section first outlines existing processes without considering computational tools. The effective development, evaluation, and deployment of AI tools will require an understanding of the manual tasks they are expected to support. This is then used to identify areas for augmentation in the *AI Method Mapping* (Section III).

This taxonomy applies to the analytic tasks in statecraft, particularly where outputs can be represented, inspected, or evaluated. This is not an exhaustive process theory of statecraft — the domain is too broad — but instead a functional decomposition of tasks that identifies appropriate methods and evaluation criteria.[19]

Our definition draws on 'practice theory' in academic International Relations, the central claim of which is that competent diplomatic performances are built on tacit, embodied knowledge that cannot be fully codified (Adler and Pouliot 2011). This work captures the limits of automation and characterizes the human work presently outside AI capabilities. Our taxonomy avoids actively codifying competence but maps the analytic tasks and space where actors can operate more or less competently. There is a growing literature in domains of practice that examines the role of digital tools, including AI, in peace mediation (Hirblinger 2022).

We propose a functional task architecture that decomposes the analytic work of statecraft into six task families by their inputs, transformations, and outputs. This captures the work that practitioners undertake to support the practices of statecraft, where teams collect information, conduct analysis, design strategies, execute plans, and evaluate outcomes. These task families are not chronologically derived, or seen as natural stages, but provide a scoped architecture for identifying recurring task families relative to

19 Participants in statecraft include sovereign states and international secretariats to legislators and third-party mediators. Negotiations can range from informal side-meetings to legislative debates and crisis diplomacy. Relevant fora vary along shared dimensions, such as value systems, transparency, formality, and epistemic traditions, and incorporate heterogeneous knowledge sources.

computational augmentation.

The task families are organized by function and can be summarized as follows:

- **Setup** tasks convert ambiguous situations into intelligible representations;
- **Research** tasks aggregate dispersed evidence into a single base;
- **Analysis** tasks convert evidence into analytic outputs like descriptions, diagnoses, or predictions;
- **Strategy** tasks convert judgments and preferences into sequenced plans;
- **Execution** tasks instantiate strategy into commitments; and
- **Monitoring** tasks ground evidence into assessments and lessons.

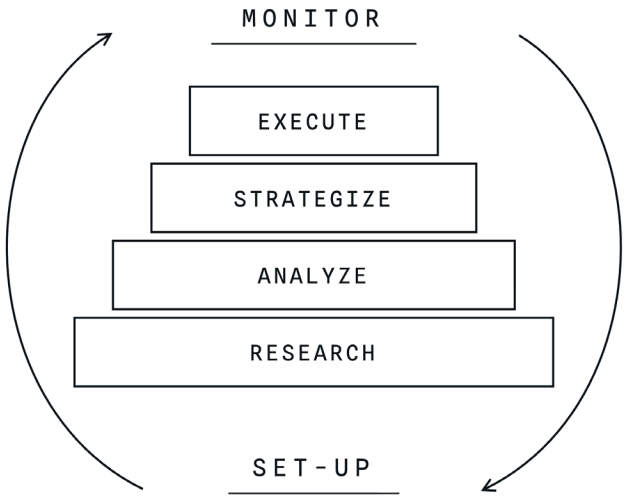
To take a concrete example: aggregating a counterpart's speeches comprises *research*, while inferring red lines from those speeches would be *analysis*. If an actor wants to test those inferred red lines, they create a *strategy*, to implement it in a live exchange would be *execution*. After coming to an agreement, seeing if commitments are honored would be *monitoring*. This allows the taxonomy to capture the iterative character of statecraft without collapsing into an undifferentiated workflow.

These six families are a functional decomposition that overlaps, iterates, and recurs. In this setting, a task is a repeatable unit of work with identifiable inputs, an applied transformation, and an output that can be inspected, used downstream, or evaluated. This taxonomy should be treated as an initial conceptual architecture subject to empirical refinement. These families are constructed so that their sub-tasks share inputs, transformations, or evaluation criteria, while tasks in different families differ on at least one of these dimensions. This is most applicable to negotiation-centered, institutional, diplomatic, and geoeconomic contexts, instead of kinetic military operations or pure political bargaining.

The six task families provide a different focus for echoes of familiar work, particularly in the intelligence cycle[20] the policy cycle, and the preparation frameworks of negotiation analysis (Raiffa et al. 2002; Lax and Sebenius 2006). Those frameworks describe sequenced workflows, whereas this paper's taxonomy enables a mapping to technical AI/ML methods and evaluation processes. This paper also extends a longer lineage of negotiation support systems: early technical efforts in the 1990s and 2000s pursued similar ambitions with contemporaneous technical tools (Kersten and Lai 2007). The architecture of the Project builds on this to include modern AI/ML techniques.

20 Direction, collection, processing, analysis, dissemination, feedback.

Figure 1: A Proposal for Standardizing Tasks in the Practices of Statecraft



Practices of Statecraft

Setup

Each process has a basic architecture: the parties, the fora, its rules, and the decision rights under which they operate. This architecture may be decided by the institutional forum or by mutual consent of parties. Parties are typically stable, but their beliefs and stated positions may shift rapidly as information and incentives change. Fora vary in mandate, membership, and procedure, with both formal and informal rules that constrain behavior and, by extension, the feasible set of options. Many of these rules are tacit and well understood by experienced teams but often remain uncodified.

Principals and their teams establish an understanding of interests, preferences, positions, and strategies, which are updated throughout the negotiation. Drawing on the extensive scholarship on negotiations, we distinguish preferences from positions:

- *Preferences* formalize interests as orderings over possible outcomes (Lake and Powell 1999; Frieden 2020). We distinguish *preferences over outcomes*[21] — what actors prefer from the set of available options — from *preferences over strategies* — what actors will do to achieve those outcomes. The same outcome preferences could generate different strategy preferences depending on beliefs about certain aspects, like feasibility, risk tolerance, and time horizons.
- *Positions* are the stated demands or proposals that parties advance. These should be seen as strategic signals —they may or may not reflect underlying preferences accurately — but they are part of a counterpart's strategy to achieve their goals (Lax and Sebenius 1986).
- *Strategies* are plans designed to achieve preferred outcomes. These are often based on beliefs about

21 Outcome preferences can be characterized as *objectives* in the decision theoretic tradition of Keeney and Raiffa, 1993.

counterpart preferences and likely trajectories of domestic ratification (Putnam 1988).[22]

Assessments of counterpart preferences are best treated as uncertain priors that will be continuously updated as new information arrives. Preliminary interest mapping might reveal where alignment is immediately plausible or infeasible. Teams must also develop their best alternatives to negotiated agreements (BATNAs), predict zones of possible agreement (ZOPA), and revisit them as needed as the negotiation progresses (Lax and Sebenius 2006).

Illustrative tasks include:

- Identify relevant parties, their positions, and assessments of their underlying preferences.
- Establish initial estimates of BATNAs, reservation values, and plausible ZOPAs.
- Detail the team's mandate and red lines.
- Identify forum rules and norms.

Research

Diplomatic action is founded on an evidentiary base — all subsequent analysis and strategy rest downstream. In strategic settings, research tasks assemble the relevant information. This eventually bounds the feasible set of agreements and defines the drivers of negotiation — such as upsides, downsides, outcomes, and risks. Teams must understand the relevant actors, the substance of the issues, the nature of the forum, and the depth of past decisions and precedents.

Currently, this work is performed by teams with varying access to resources and contextual information. Teams may have differing access to tools or tacit knowledge distributed across staff. Our interviews with practitioners underscore the scale of this informational burden: Effective negotiation preparation requires assimilating large volumes of material under tight time constraints.[23]

This research can be updated throughout the course of a diplomatic negotiation. Changes to the evidence base impact downstream tasks, with analyses and strategies changing based on new information on-the-ground. Illustrative research tasks:

- **Collection**: retrieving documents, data, and signals from primary sources and secondary commentary, such as legal texts, prior agreements, intelligence products, and expert consultations.
- **Validation**: assessing source reliability and cross-checking claims to guard against misinformation and unverifiable assertions.
- **Structuring**: organizing raw material into comments on issues, parties, and their positions.
- **Synthesis**: generating summary briefs, position maps, and background notes that can be inspected, updated, and reused by the wider team.

Analysis

Analysis tasks occur continuously, converting raw information into coherent decision-relevant formats. Teams move from description to diagnosis and prediction, explaining drivers and implications. Analyses cover

22 In the formal game-theoretic sense, a strategy is a complete contingent plan specifying actions at every possible decision point (von Neumann and Morgenstern 1944); in practice, diplomatic strategies are rarely so complete, but the emphasis on anticipating counterpart responses remains central.

23 Synthesized findings from our initial round of interviews with senior practitioners are forthcoming in 2026.

both quantitative and qualitative methodologies across multiple layers of agency.

Multiple analyses will likely occur simultaneously and are rarely reducible to a single lens. Outcomes reflect more than rational optimization, but encompass routines, procedures, and internal bargaining (Allison and Zelikow 1999; North 1990; March and Olsen 1983). Strong analyses, therefore, both separate and synthesize perspectives from different actors.

On quantitative issues, analysis may involve statistical modeling to estimate future developments, quantifying the likely trajectories, payoffs, or impacts that might unfold from a given situation. Scenario analysis through decision trees can outline plausible futures and clarify trade-offs for human teams. On qualitative issues, analysis often centers on applied logical reasoning, drawing on legal justification or historic precedent. Teams can identify procedural constraints within a given forum, typical coalition patterns, and salient precedents that shape both the possible and the likely options.

Outputs of analysis can be seen as processed models of the negotiation environment, looking at rearranging some constituent part, or illustrating some aspects of the situation in their entirety. Illustrative tasks include:

- Conducting descriptive, diagnostic, predictive, and prescriptive analysis.
- Undertaking legal, economic, strategic, causal, or political analysis.
- Identifying affected channels (trade, sanctions, finance, energy, aid) and estimating costs/benefits and distributional impacts.
- Interpreting text-based data (speeches and texts) to infer priorities and red lines or identify key narratives shaping perceptions or domestic politics.
- Analyzing coalitions and central actors, with plausible realignments under different issue packages.

Strategy

Strategy ties preferences and constraints to a sequenced plan of actions, to achieve "the alignment of potentially unlimited aspirations with necessarily limited capabilities" (Gaddis 2018). In our framing, strategy converts research and analysis tasks into plans to execute on a team's preferences. Strategies leverage conceptual frameworks, like negotiation theory and bargaining theory, to outline packages and shape the developing bargaining dynamics.

Game theory and bargaining models can contribute to strategies by exploring the range of credible bargains and theorizing mechanism design for enforcement (Schelling 1960; Fearon 1995). Decision-analytic tools can structure trade-offs and uncertainty across this space, clarifying priorities and contingencies.

Illustrative tasks include:

- Translate preferences and constraints into concrete objectives.
- Design creative strategies that convert objectives into a series of executable moves.
- Construct potential packages of issues that expand or exploit the ZOPA, specifying sequencing of offers and signals.
- Anticipate counterpart responses, scenarios, and potential failures.
- Outline concrete success metrics.

Execution

Execution involves an ongoing interplay between relationships at the table and internal team tactics, with dynamic adjustments occurring as information evolves. The nature of execution varies widely across settings, but these processes are often conducted under time pressure and uncertainty, featuring persistent private information and knowledge gaps (Fearon 1995; Lake and Powell 1999). Teams must translate strategies into offers and signals while managing multiple audiences, from counterparts at the table to domestic principals and public observers.

Diplomatic negotiations can break down in execution even when research, analysis, and strategy are sound. Effective execution requires balancing strategic discipline with emotional intelligence, integrating an understanding of individual motivations with institutional priorities. Conflicting instructions, siloed information, and competing bureaucratic interests can derail execution (Allison and Zelikow 1999; Halperin et al. 1974). Additionally, interpersonal dynamics between negotiating representatives can expand or constrain the space of possible outcomes.

Illustrative tasks:

- Translate strategies into offers and communications.
- Track positions and commitments across teams, channels and time.
- Propose what information to share or hide at each point.
- Recommend potential offers.
- Agree on monitoring and enforcement methods and milestones.

Monitoring

Once an agreement has been reached, monitoring activities track implementation, creating an evidentiary base for future engagements. This includes synthesizing a broad range of information to assess both compliance and outcomes. Compliance, the honoring of agreed commitments, can be determined through mechanisms like inspections or open-source intelligence.

Well-designed monitoring can reduce uncertainty and distinguish intentional noncompliance from capacity shortfalls or ambiguities.[24] Clear indicators, and data-sharing procedures will be essential for verification of compliance. Without explicit reconciliation protocols, the durability of commitments can be undermined on technical, political, and institutional grounds.

Monitoring and enforcement provide the raw material for institutional learning. Systematic evaluation of implementation should feed back into future negotiations. Institutions that encode these lessons can build resilience over time, reducing the likelihood of repeating past errors. In this sense, monitoring and enforcement function as preparation for future negotiations. These processes generate institutional memory that reshapes the feasible design space for subsequent processes.

Illustrative tasks include:

- Track implementation milestones against timelines and specified benchmarks.
- Synthesize evidence into assessments.

24 In repeated games, automatic sanctions and snapback provisions can underpin cooperation by altering the long-run incentives of parties (Fearon 1998; Axelrod 1984).

- Distinguish intentional and inadvertent noncompliance.
- Assess whether outcomes are achieving underlying objectives.
- Document precedents and design lessons for future scoping and strategy.
- Trigger strategy revision when needed.

Conceptualizing Improvements to Existing Tasks

Improvements can broadly take two forms: process and outcome improvements. Process improvements can be characterized by different levels of agency (individual, team, institution, and forum), while outcome improvements require some capacity to provide a holistic overview on a given outcome.

As outlined in Section I, we can broadly characterize process improvements along speed, quality, range, and novelty. A tool might be able to expand the range of options considered, without any of them being novel (perhaps by surfacing known options a team lacked time to enumerate), and can produce novelty without meaningfully expanding the range of options considered, through providing a single unconventional proposal.

The challenges of building effective computational tools identified in Section I (representation, inference, and evaluation) affect these improvement dimensions differently:

1. ***Speed* gains are least constrained.** If the same task can be completed and verified faster without sacrificing quality or safety, that is valuable. Research and synthesis tasks offer the clearest speed gains because they operate on documented materials with specifiable processes.
2. ***Quality* improvements are partially constrained.** Quality in research, such as coverage or accuracy, can be measured against some standards. Analytic quality is harder to assess without ground truth — such as inferred counterpart preferences. Augmented analysis should improve process quality while acknowledging that substantive quality may remain unknowable.
3. ***Range* is predominantly constrained by representation issues.** We can reliably expand which options are considered only in structures we can represent. Computational tools that can enumerate combinations or explore scenarios depend on accurate representations. Novel solutions may emerge from LLMs, but the methods of generation cannot ensure their validity.
4. ***Novelty* is constrained by representation and evaluation challenges.** Novel options can be generated through recombination, departures from precedent, and unconventional packages. Determining whether novel options are genuinely better requires resolving the evaluation problem: defining "better" for whom, on what dimensions, and with what legitimacy.

Section III continues by mapping computational methods to each task family.

Box 1. An Illustrative Human-Only Diplomatic Negotiation Process

Setup. Define the problem, objectives, and constraints. The team fixes the mandate, red lines, and acceptable fora. This might include forum parameters (decision rule, timetable, procedural rules, confidentiality) are recorded.

Research. Analysts compile the evidentiary base: legal texts and precedents; counterpart positions and domestic constraints; economic, political, and security context; stakeholder maps; and recent signals. Sources are validated and logged. Outputs include briefs or position maps.

Analysis. Implications from the evidence: feasible sets under forum and domestic constraints, likely coalitions, anticipated responses, legal viability, and operational risks. Quantitative estimates and structured scenario analysis bound uncertainty.

Strategy. Negotiators compose an option set and a sequencing plan: opening bid, concession ladders, issue linkage, assurances, and fallbacks. Within this, each option carries success conditions, monitoring indicators, and pre-approved talking points. A negotiation playbook aligns roles (lead, legal, policy, technical, comms) and specifies escalation paths for on-the-spot decisions.

Execution. The team executes the plan across bilateral and multilateral sessions, using signaling and information release. Offers and counteroffers are evaluated against red lines and the reservation point. Deviations require rapid internal response. Provisional understandings are converted to text and checked for legal compliance while definitions and enforcement clauses are reconciled.

Monitoring. After agreement, the team activates follow-up mechanisms (timelines, verification, dispute resolution) and assigns owners for each obligation. Communications are coordinated. Compliance and outcomes are tracked against predefined indicators.

III. AI Method Mapping

This section connects each task family to types of AI/ML tools that could add value under realistic constraints. We identify three broad categories of AI tools: knowledge systems, predictive models, and decision systems. This section proposes AI/ML methods to augment those tasks, alongside expected outputs and potential metrics for process and outcome assessments.

Any deployment of AI systems in real diplomatic settings will be shaped by a range of constraints not wholly considered in this initial conceptual architecture. Statecraft is built across a range of classifications, disclosures, procurement practices, networks, and norms. For example, real-time transcription may be prohibited, or information networks may be physically prevented from information aggregation. Task-level augmentation, therefore, must be built around a range of *de facto* and *de jure* limitations. The institutions themselves, however, will need to adapt to these constraints to remain competitive as the information and technical environments evolve.

AI/ML Methods

We distinguish three broad types of computational method that occur commonly across applications in strategic domains, like statecraft.[25] These are knowledge systems, predictive and inferential models, and policy and decision systems — plus a fourth, cross-cutting construct: world models, which integrate the other three and supply the representational substrate they operate on.

Knowledge systems (able to structure, store, and search information).

These are systems whose primary function is to interact with information. LLMs are powerful tools for processing and generating natural language, but their outputs reflect distributional patterns in training data. One of their key values in diplomatic settings lies in retrieving, summarizing, and synthesizing data. This is particularly true when the primary challenge is handling large volumes of natural language material at speed without sacrificing source fidelity.

Predictive and inferential models (supervised and unsupervised learning)

These are models whose primary function is to estimate hidden variables or forecast outcomes from text and structured signals. They take structured inputs like numbers, categories, or text, and produce estimates like probabilities, classifications, or scores by learning quantitative relationships from labeled examples (supervised learning) or discover latent structure in unlabeled data (unsupervised learning).[26] In strategic contexts, states could be clustered into coalitions based on voting behavior or latent factors driving alignment patterns could be extracted from voting records.

Policy and decision systems (RL and multi-agent RL)

Systems whose primary function is to find optimized actions under uncertainty, often with explicit objectives

25 Two clarifying notes prevent category confusion. First, many knowledge systems are powered by large language models that were pretrained with self-supervision and often refined with preference optimization. What makes them distinct here is the grounding function in data as opposed to the learning paradigm. Second, world models frequently incorporate RL components, but they are separated because their defining feature is explicit counterfactual evaluation and component integration, rather than policy learning alone.

26 In practice, this class includes familiar tools (linear and generalized linear models, decision trees, random forests, gradient-boosted machines) as well as deep encoders that map text or graph structures into continuous vector spaces. Their role in strategic settings is to provide calibrated indicators and structured summaries rather than final decisions: risk scores, coalition maps, and estimates of key quantities of interest that feed into human and algorithmic reasoning downstream.

and constraints. Reinforcement learning (RL) formalizes sequential decision-making as an agent interacting with an environment, receiving observations, taking actions, and accruing rewards. For diplomatic settings, RL and MARL are relevant in two distinct roles: forward policy search, in which agents stress-test strategies inside a formal model, and inverse inference (IRL), which recovers approximate reward functions from observed actions. The roles fail differently — a policy search inherits every defect of its simulator, while inverse inference inherits the problems described in Section I. Alongside RL, evolutionary and population-based search methods explore strategy spaces without requiring differentiable objectives — a fit for the discrete, combinatorial character of issue packages.

World models

To be leveraged to their fullest extent in statecraft, these method classes require some form of representation of the world. In these settings, a world model is a structured, dynamic representation of the environment. These representations are neither pure predictors nor pure RL agents but instead simulate the world and can generate defensible sequences of events that would unfold if certain conditions held and certain actions were taken. Such simulators can provide controlled environments to explore "what-if" questions and to test the robustness of strategies. They would be most useful if explicitly derived from data and bound by the constraints of political realities.

As discussed in Section I, no unified model of world politics is currently feasible — the scale of simulation would be astronomical. Instead, partial, bounded representations that are valid in their specific subdomain, could be incredibly useful. At a minimum, such a model would need to encode:

- **States**. In computer science literature, these are "states of the world" as opposed to nation states. These "states" encompass latent and observed variables describing actors, capabilities, preferences, beliefs, issues, fora, and commitments. This includes domestic constraints and institutional rules, not just dyadic preferences.
- **Transition dynamics**. How states evolve over time as events occur. This involves simulating the impacts of events — in statecraft this might include how offers, sanctions, resolutions, or crises alter the state of the world.
- **Observation model.** How text, numbers, and other signals (like news reports, speeches, votes, economic data) map onto state variables.[27] This connects raw data to the conceptual quantities that negotiators care about.
- **Reward models**. How actors benefit from certain states.[28] Do they care about things like security, economic welfare, status, or normative recognition?

In practice, the first step toward constructing a valid world model is a consistent way to represent knowledge. For research and analysis tasks, this requires accurate information encoding and processing. Models involved in strategies, however, need to ensure downstream methods, like predictive models, RL agents, or simulators, operate within realistic feasibility constraints. This implies any world model is not a fixed database but can support a living representation that responds to new events and information in the system. 'Online learning', where models continuously learn and update in response to data, could be a technical means to this end.

27 Potential frameworks to execute on this include tools like Apache Jena RDF stores, graph-neural networks, or JEPA models for scoped domains. Scoped tasks might include predicting alliance formation or propagating event shocks across issues.

28 Inferring a reward function requires an inverse-reinforcement-learning or imitation-learning setup with behavioral data and an environment model.

Box 2: Illustrative AI-enabled COP negotiation

Imagine a delegation preparing for a Conference of the Parties (COP) session on loss-and-damage financing. How might the Task Taxonomy and AI Method Mapping help a team conduct their work?

Setup. The team specifies the negotiation environment. This covers different interests (for example, small island states, major emitters, emerging economies), the forum (in this case UNFCCC COP), and best estimates of initial positions drawn from experience and a first interpretation of national submissions. This information populates the knowledge base, potentially as a graph connecting parties to their positions.

Research. Building from this foundation, AI tools could comprehensively develop the knowledge foundation. Retrieval-augmented language models process the corpus of prior COP decisions, public national adaptation plans, media coverage, and the academic literature on climate finance to ensure a comprehensive knowledge foundation. An AI system surfaces precedents (such as the Green Climate Fund's governance structure), identifies gaps (no prior decision addresses a specific liability question), and flags contested claims (disputed estimates of climate-attributable losses). From this, the human team can review cited summaries of the information environment, with team analysts verifying, contributing to, and correcting the database.

Analyze. Material from the research stage is processed into varying analyses. This includes replications of economic forecasts, estimates of financing levels, and analyses of historical concession patterns. Clustering algorithms identify latent coalitions that cross traditional bloc lines, while influence diagrams map how changes in one issue could affect positions on connected issues. From this stage, the team has a quantified estimate of the zone of possible agreement and a sense of the interconnections between parties and issues.

Strategize. Within a simulation sandbox, RL agents explore package proposals: varying financing commitments, thresholds, and governance arrangements to see how parties might respond to each arrangement. The system identifies proposals that fall within the estimated ZOPA and stress-tests them against simulated counterpart responses. One candidate package proves robust across a range of plausible counterpart preferences and the team stress-tests the assumptions and preliminarily adopts it as a primary option with two fallback variants.

Execute. During the negotiation, real-time transcription and entity extraction update the knowledge base as parties make statements that reveal more about their preferences. An AI system flags when a key swing party's language shifts from prior positions, alerting the team to a potential opening. Draft text is assessed against the team's known red lines in real time.

Monitor. After an agreement, the system tracks implementation progress, with specified indicators like disbursement rates, eligibility determinations, and compliance with reporting requirements. Deviations to agreed options can trigger alerts to all parties, with systematic evaluation feeding lessons into preparation for the next session.

Method Mapping by Task Family

There are many types of tasks within the practices of statecraft; this mapping explores which

augmentation methods are (i) currently feasible, (ii) near-term research targets, and (iii) currently speculative. It is important to note that this is a series of hypotheses about capabilities and that there is a pervasive gap in the current evaluation practices that needs broad attention (Pozniak and Sania 2026).

Setup

Methods:	World models and simulators built from knowledge graphs and ontologies.
Key challenge:	Representation.

In a minimal form, setup tasks might simply store the actors and their positions in a simple database that is accessible to team members. Such structured knowledge representations could improve downstream task performance. If team members can comprehensively query a knowledge base on parties or issues, then subsequent outputs could be both more accurate and complete, while also being produced more quickly. The utility might degrade with active contestation of institutional rules, however.

More sophisticated models might capture this through graph-based knowledge maps, with analyses conducted over relationships between nodes. The frontier research agenda may be a structure that supports small, simulated dynamics of the interactions between parties and issues. Accurately capturing states could support structuring information and analysis, while transition dynamics could support strategies or counterfactual analyses.

Research

Methods:	Retrieval-augmented LLMs, web-browsing agents, database collection.
Key challenge:	Inference.

Negotiation teams must gather large volumes of primary and secondary sources and store them to support analysis. Retrieval-augmented LLMs and web-browsing agents can search vast databases and assemble information. This might include building counterpart biographies or compiling a list of relevant resolutions. This frees negotiators for analytic tasks requiring synthesis and judgment.

The capability limits for retrieval-augmented LLMs are currently being evaluated across a range of industries, with the potential for increasing the coverage and speed of research tasks becoming apparent. For a defined domain, a corpus must be curated and the retrieval grounded in citations, presenting source extracts as opposed to text generation. Evaluations here should be focused on precision, recall, and coverage in the data retrieval architecture.

When facing data scarcity (the inference problem), incomplete records abound; as such, source validation and gap detection are essential research functions that lend themselves to existing capabilities of LLMs with tool use. Such tools could improve the quality of analysis and the range of defensible positions by stress-testing assumptions and identifying omissions or provenance issues. They could not, however, detect omissions arising from evidence never recorded.

Analyze

Methods:	Supervised prediction models, unsupervised and representational learning models, tool-augmented agents.
Key challenges:	Representation and inference.

For analysis tools to be useful, the historical data would need to be sufficiently rich and the coalition dynamics sufficiently stable to ensure predictive power on new issues. Supervised machine learning models could project relevant quantities or estimate event probabilities with greater accuracy than human analysts. Continuous quantities, such as expected costs, or event probabilities, such as noncompliance or defection likelihood, could be predicted using a range of inputs like past voting behavior or economic indicators. Accurate prediction could then improve the quality and novelty of proposed strategies.

Probabilistic estimates of relevant quantities, like the zone of possible agreement (ZOPA) or enforcement payoffs could in principle be better calibrated than unaided expert estimates. However, because these quantities are typically latent or counterfactual (rather than observed), these estimates would depend on proxies and rules meaning that outputs should be treated as conjecture rather than ground truth (Fearon 1995; Koremenos et al. 2003).

Natural language processing (NLP) could complement supervised models through text-based position signals extracted from speeches or statements. This would test whether position or sentiment extraction could be informative. This could be tested through historical negotiation transcripts with known eventual outcomes to assess in-sample fit and out-of-sample prediction.

Unsupervised methods (clustering, graph-based learning, deep encoders) can identify latent coalitions and cross-cutting alignments invisible in raw data. Where preferences are partially hidden, inverse reinforcement learning (IRL) may infer potential reward structures from historical action sequences. As Section I established, however, such estimates must be treated as hypotheses requiring qualitative expertise as opposed to an *ex ante* verifiable ground truth.

Ensuring robustness requires disciplined validation and close integration with domain expertise. Evaluation should proceed through internal consistency checks, uncertainty quantification, and out-of-sample validation. These criteria are demanding but feasible because analysis has been separated from the downstream strategic judgments it informs.

The representation problem also impacts the range of available techniques, as certain methods remain unable to capture the whole system of world politics. Partial models of actors, issues, or institutional rules are tractable, but whole-system models capturing how institutions, preferences, and actors co-evolve remain beyond current formalizations. AI-assisted analysis should be understood as operating within assumed institutional structures and balancing the symmetry between description and prescription.

Strategize

Methods:	Reinforcement learning (RL), multi-agent RL, evolutionary search.
Key challenges:	Representation, inference, evaluation.

Strategy-generation tasks are where the decomposition logic is most consequential. The *evaluation* problem requires resolving a range of tradeoffs between certain normative values. Assessing legitimacy across audiences and judging ratification feasibility involve empirical components that tools can help estimate, and political choices that AI systems should not make. Augmentation in strategy generation focuses on expanding and stress-testing the option space while leaving the selection of strategies and their respective trade-offs to human decisionmakers.

RL and MARL systems have matched or exceeded expert performance in complex games, but in world politics, these methods may be most valuable for converting stylized bargaining problems into strategy sandboxes as opposed to trying to 'solve' a situation (Mnih et al. 2015; Silver et al. 2018; Vinyals et al. 2019; Meta Fundamental AI Research Diplomacy Team (FAIR)† et al. 2022). When given actors, potential actions, rules, and payoffs, agents could explore the space of possible strategies and counterstrategies, generate novel combinations of mechanisms, and test their robustness in simulation.

The success of such strategy simulations would be in the range and evaluated quality of these options, particularly when compared to conventional brainstorming and tabletop exercises. Critically, the institutional constraints and behavioral models in simulation must be sufficiently realistic for the options to be feasible for practitioners.

Additionally, human-proposed strategies could be stress-tested against models that identify vulnerabilities or simulate public responses that strategists might miss. As opposed to exploring the range of potential strategies, agent-based simulation systems could help teams explore the potential responses and consider adapting responses. Such simulation techniques are beginning to emerge in private sector research to assess communications, advertising, and corporate strategy problems. Predicting a group's response to a given strategy is easier than constructing an optimal novel one as evaluating a fixed candidate is a narrower problem than open-ended generation.

Execute

Methods:	Automated speech recognition, translation, tool-augmented LLMs, sentiment analysis.
Key challenges*:*	Inference.

Execution translates strategy into real-time action under tight temporal and informational constraints. At a basic level, real-time transcription, entity extraction, and commitment tracking during negotiations could reduce the rate of internal inconsistencies and untracked commitments relative to manual notetaking. This is immediately testable in simulated negotiation environments and concerns observable process metrics.

During negotiations, language models could instantly check the consistency of public and private statements against structured information. This might include precedent, mandates, or prior positions, both within teams and between counterparts. This could improve both the quality of communications and understanding across participants.

Beyond the generation and comparison of texts is the capability boundary for AI systems to recognize strategic misrepresentation and the subtleties of diplomatic language. It will be critical to understand the boundary between tracking *what was said* and *what was meant* by different parties. This will require both the production of tools and diplomatic datasets to compare stated preferences, strategic misrepresentation, and ground truths.

These tools should not replace in-person diplomatic presence; these are not 'one-stop-shops' that can take an issue and output a verifiable plan. These tools must support human decisions by providing dynamic, informed assessments with multiple options outlined. Creativity in execution is about expanding the space of considered options within an existing strategy.

Monitor

Methods:	Retrieval-centric language systems, supervised prediction, time-series models, game-theoretic simulation.
Key challenge:	Inference.

Once an action has been taken or an agreement executed, monitoring closes the loop through tracking implementation and feeding the evidence back into the collective research base. Timescales vary from real-time verification to decades-long outcome assessments. Automating parts of the monitoring process would lessen the administrative burden of tracking hundreds of specific obligations, timelines, and conditionalities.

Automated commitment tracking could be built with strong data aggregation in settings with quantifiable, verifiable commitment clauses. Such processes could increase the coverage of teams under limited resources or decrease the response time in the event of a breach or a new crisis. Such systems could leverage a range of classic machine learning prediction tools, such as time-series anomaly detection. These systems could be evaluated as classic prediction models, checking false-positives and false-negatives of flags, but could be undermined by a lack of ground truth in both training and testing. Agreement breaches may be from cheating, a lack of institutional capacity, or misinterpretation. These systems, no matter how automatable, should keep their information-sharing augmentation function.

AI augmentation here can make the feedback loop tighter, more systematic, and less dependent on individual memory. Novel methods for enforcement or monitoring are more likely to emerge from institutional and treaty design decisions, rather than in monitoring techniques.

The table below summarizes this analysis and provides an operational framing. It consolidates the method mapping and presents both the evaluability and maturity of the given families. This helps inform practitioners on which methods have the most potential, as well as targeting a research agenda for technologists on the most immature areas. The initial scope of augmentation should be aligned with existing processes that can be reliably verified. This architecture is a set of hypotheses as to the potential feasibility of augmentation, as opposed to a catalogue of finished tools.

Table 1: Consolidated Method Mapping by Task Family

Family	Outputs and Methods	Challenges	Evaluability Tier	Maturity
Setup	World model or knowledge graphs, entity extraction	*Representation*	(ii) expert-auditable	Deployable
Research	Structured evidence base through RAG LLMs	*Inference*	(i) externally checkable	Deployable
Analyze	Supervised/ unsupervised models conducting analyses	*Representation* + *inference*	(i)–(ii) mixed	Near-term
Strategize	RL/MARL in simulation, evolutionary search to provide validated options	*Evaluation*: candidates can be tested not scored	(ii)–(iii)	Speculative / research frontier
Execute	Transcription, tracking, option generation, and other real-time support	*Inference*: signals designed to mislead	(i) for process metrics; (iii) for meaning	Deployable (process layer)
Monitor	Retrieval, classification, anomaly detection	*Inference*: intent uncertain post-deviation	(i) for indicators; (ii) for intent	Deployable in scoped domains

Incorporating a task-based decomposition within a complex system will require explicit points of control where distinct steps meet each other, or human interaction. At a minimum, user interfaces should note the upstream provenance of any information used in a downstream claim. Ensuring that humans are informed on how the pieces fit together will help keep decision support from slipping into automation.

Open Questions and Path Ahead

Overview

The twenty-first century will be shaped by whether states can resolve their differences without catastrophic conflict. The practices of world politics and statecraft are straining under the weight of accelerating complexity and geoeconomic pressure. Artificial intelligence will transform these processes whether we act deliberately or not. Whether that transformation will be disciplined or reckless, whether it will preserve human judgment or erode it, whether it will serve the pursuit of durable peace or accelerate the dynamics of miscalculation depends on the steps we take now.

Scoped, modular augmentation must respect the structural constraints of diplomatic environments while capturing genuine gains in speed, quality, range, and novelty. End-to-end automation is infeasible — the representation, inference, and evaluation problems are constitutive features of strategic interaction that serious scholarship has long theorized. Blanket skepticism of AI's utility is also not warranted: Within well-defined boundaries, computational tools can accelerate research, sharpen analysis, expand strategic

imagination, and strengthen institutional memory.

This whitepaper establishes a positive agenda for a new line of ambitious interdisciplinary research to resolve the hardest problems in implementing AI systems in the international system. The Task Taxonomy and the AI Method Mapping are an approach to the challenge of disciplined AI augmentation. This whitepaper outlines a scaffold for disciplined progress. It provides practitioners with a structured way to evaluate which tools might help and which claims are overreach. It provides researchers with a demand-side view of where capabilities are needed and where they already exist. It provides policymakers with a basis for governing AI deployment in high-stakes domains.

Open Questions

Major questions remain open, and their answers will determine the success of AI augmentation in the practices of statecraft, particularly around the future of diplomatic institutions.

What decisions can we legitimately delegate to computational systems?

Current proposals, like having LLMs express uncertainty quantitatively or track source provenance, mitigate but do not solve this problem. What institutional arrangements, training regimes, or system designs would allow practitioners to appropriately calibrate trust?

How does the geostrategic environment change when multiple parties use AI systems?

Error clustering introduces a systemic risk, but its dynamics are currently poorly understood (Daníelsson et al. 2022). If major powers adopt similar AI tools, will their analyses converge in ways that reduce bargaining space? Will correlated blind spots produce coordinated misperceptions? Will adversaries exploit known system limitations? These questions require empirical investigation that does not yet exist.

How do we combine well-delegated (human and machine) tasks into larger, complex strategies?

This paper focuses on process improvements: speed, quality, range, and novelty in task performance, but each task needs combining to a larger strategy. Whether better processes yield better outcomes, fewer conflicts, or more legitimate resolutions is an empirical question we cannot currently answer. The relationship between the process quality of each task, and the quality of the whole outcome, is not straightforward in complex adaptive systems.

What becomes the comparative advantage of humans and institutions as AI capabilities advance?

Current AI capabilities are limited in several facets, as explored in this paper, but systems will continue to improve. AI tools will continue to increase performance in tasks currently reserved for humans. Designing effective institutional arrangements to ensure meaningful human authority is as much a governance question as a technical one. This requires attention now, before path dependencies lock in.

Conclusion

The Project on Computational Statecraft is committed to pursuing this research agenda: building and validating tools, establishing evaluation frameworks, developing institutional arrangements that center human authority, and contributing to a field that will shape how humanity manages its most consequential disagreements. This paper sets the initial architecture to inform an empirical and theoretical research program.

Near-term work will focus on empirical and infrastructural developments — investigating the suitability of existing model architectures for various tasks within the taxonomy. Extensive interviews with high-level practitioners are informing future design while diligent technical and theoretical work will also be necessary to carve out a new theory of applied model evaluations in high-stakes settings.

AlphaGo's move 37 revealed a possibility that centuries of human expertise had missed. We believe that disciplined AI augmentation can surface possibilities in world politics that current practice overlooks: institutional designs that expand zones of agreement, enforcement mechanisms that prove more robust, framings that make previously blocked proposals legible to different audiences. But AlphaGo's move was played within fixed rules by a system with clear objectives. To make meaningful progress in the practice of statecraft, we must face the challenges that come with computational augmentation.

The practitioners of statecraft should be neither credulous nor dismissive of AI. They must demand evidence, insist on transparency, and retain authority over consequential decisions. The challenge today is to ensure that we are building tools adequate to the complexity of a changing world.

Bibliography

Abbott, Kenneth W., Robert O. Keohane, Andrew Moravcsik, Anne-Marie Slaughter, and Duncan Snidal. 2000. "The Concept of Legalization." *International Organization* 54 (3): 401–19. https://doi.org/10.1162/002081800551271.

Adler, Emanuel, and Vincent Pouliot. 2011. "International Practices." *International Theory* 3 (1): 1–36. https://doi.org/10.1017/S175297191000031X.

Allen, Gregory C., and Taniel Chan. 2017. "Artificial Intelligence and National Security." Belfer Center for Science and International Affairs, Harvard Kennedy School. Preprint, July. https://www.belfercenter.org/publication/artificial-intelligence-and-national-security.

Allison, Graham T., and Philip Zelikow. 1999. *Essence of Decision: Explaining the Cuban Missile Crisis*. Second edition. Longman. http://swbplus.bsz-bw.de/bsz083131086inh.htm.

Ancheva, Slavina, Charles Pozniak, and Sameera Salari. "Perspectives from Practice: Interviews with Senior Practitioners on AI and Diplomatic Negotiations." Belfer Center for Science and International Affairs, Harvard Kennedy School, manuscript in preparation, 2026.

Axelrod, Robert. 1984. *The Evolution of Cooperation: Revised Edition*. Basic Books.

Aydoğan, Reyhan, Katsuhide Fujita, Tim Baarslag, Catholijn M. Jonker, and Takayuki Ito. 2021. "ANAC 2017: Repeated Multilateral Negotiation League." In *Advances in Automated Negotiations*, edited by Takayuki Ito, Minjie Zhang, and Reyhan Aydoğan. Springer. https://doi.org/10.1007/978-981-15-5869-6_7.

Baarslag, Tim, Reyhan Aydoğan, Koen V. Hindriks, Katsuhide Fujita, Takayuki Ito, and Catholijn M. Jonker. 2015. "The Automated Negotiating Agents Competition, 2010–2015." *AI Magazine* 36 (4): 115–18. https://doi.org/10.1609/aimag.v36i4.2609.

Bainbridge, Lisanne. 1983. "Ironies of Automation." *Automatica* 19 (6): 775–79. https://doi.org/10.1016/0005-1098(83)90046-8.

Baldwin, Carliss Y., and Kim B. Clark. 2000. *Design Rules, Vol. 1: The Power of Modularity*. Mit Pr.

Baldwin, David A. 1980. "Interdependence and Power: A Conceptual Analysis." *International Organization* 34 (4): 471–506.

Barnes, Matt, Matthew Abueg, Oliver F. Lange, et al. 2024. "Massively Scalable Inverse Reinforcement Learning in Google Maps." arXiv:2305.11290. Preprint, arXiv, March 5. https://doi.org/10.48550/arXiv.2305.11290.

Barnett, Michael, and Martha Finnemore. 2004. *Rules for the World: International Organizations in Global Politics*. Cornell University Press.

Bellman, Richard. 1957. *Dynamic Programming*. Princeton University Press.

Boussioux, Leonard, Jacqueline N. Lane, Miaomiao Zhang, Vladimir Jacimovic, and Karim R. Lakhani. 2024. "The Crowdless Future? Generative AI and Creative Problem Solving." SSRN Scholarly Paper No. 4533642. Social Science Research Network, July 1. https://doi.org/10.2139/ssrn.4533642.

Brown, Tom, Benjamin Mann, Nick Ryder, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–901.

Bruderlein, Claude. 2026. "Report of the AI Negotiation Forum 11 January 2026." SSRN Scholarly Paper No. 6925859. Social Science Research Network, February 18. https://papers.ssrn.com/abstract=6925859.

Brusoni, Stefano, Andrea Prencipe, and Keith Pavitt. 2001. “Knowledge Specialization, Organizational Coupling, and the Boundaries of the Firm: Why Do Firms Know More Than They Make?” *Administrative Science Quarterly* 46 (4): 597–621. https://doi.org/10.2307/3094825.

Busoniu, Lucian, Robert Babuska, and Bart De Schutter. 2008. “A Comprehensive Survey of Multiagent Reinforcement Learning.” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38 (2): 156–72. https://doi.org/10.1109/TSMCC.2007.913919.

Chalkiadakis, Georgios, Edith Elkind, and Michael Wooldridge. 2022. *Computational Aspects of Cooperative Game Theory*. Springer Nature.

Chayes, Abram, and Antonia Handler Chayes. 1995. *The New Sovereignty: Compliance with International Regulatory Agreements*. Harvard University Press.

Clausewitz, Carl von. 1832. *On War*. Princeton University Press. https://muse.jhu.edu/pub/267/edited_volume/book/64878.

Clement, Michael B., and Senyo Y. Tse. 2005. “Financial Analyst Characteristics and Herding Behavior in Forecasting.” *The Journal of Finance* 60 (1): 307–41. https://doi.org/10.1111/j.1540-6261.2005.00731.x.

Coase, R. H. 1937. “The Nature of the Firm.” *Economica* 4 (16): 386–405. https://doi.org/10.1111/j.1468-0335.1937.tb00002.x.

Daníelsson, Jón, Robert Macrae, and Andreas Uthemann. 2022. “Artificial Intelligence and Systemic Risk.” *Journal of Banking & Finance* 140 (July): 106290. https://doi.org/10.1016/j.jbankfin.2021.106290.

Davis, Randall, Howard Shrobe, and Peter Szolovits. 1993. “What Is a Knowledge Representation?” *AI Magazine* 14 (1): 17–17. https://doi.org/10.1609/aimag.v14i1.1029.

Dell’Acqua, Fabrizio, Edward McFowland III, Ethan R. Mollick, et al. 2023. “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.” SSRN Scholarly Paper No. 4573321. Social Science Research Network, September 15. https://doi.org/10.2139/ssrn.4573321.

Endsley, Mica R., and Esin O. Kiris. 1995. “The Out-of-the-Loop Performance Problem and Level of Control in Automation.” *Human Factors* 37 (2): 381–94. https://doi.org/10.1518/001872095779064555.

Fearon, James D. 1995. “Rationalist Explanations for War.” *International Organization* 49 (3): 379–414.

Fearon, James D. 1998. “Bargaining, Enforcement, and International Cooperation.” *International Organization* 52 (2): 269–305.

Fortna, Virginia Page. 2004. *Peace Time: Cease-Fire Agreements and the Durability of Peace*. Princeton University Press.

Frieden, Jeffry A. 2020. “Chapter Two. Actors and Preferences in International Relations.” In *Strategic Choice and International Relations*, edited by David A. Lake and Robert Powell. Princeton University Press. https://www.degruyterbrill.com/document/doi/10.1515/9780691213095-003/html?lang=en&srsltid=AfmBOoqGrHZyX-YN1nkvqwVUfVwne-ZMsDuXF0uyJF8X5wmCzFERTV-x.

Gaddis, John Lewis. 2018. *On Grand Strategy*. Penguin Books.

Goddard, Stacie E., Paul K. MacDonald, and Daniel H. Nexon. 2019. “Repertoires of Statecraft: Instruments and Logics of Power Politics.” *International Relations* 33 (2): 304–21. https://doi.org/10.1177/0047117819834625.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. “Cooperative Inverse Reinforcement Learning.” arXiv:1606.03137. Preprint, arXiv, June. https://doi.org/10.48550/arXiv.1606.03137.

Halperin, Morton H., Arnold Kanter, and Priscilla Clapp. 1974. *Bureaucratic Politics and Foreign Policy*. Brookings Institution Press.

Handel, Michael I. 1981. *The Diplomacy of Surprise, Hitler, Nixon, Sadat*. Harvard University Press.

Harsanyi, John C. 1967. "Games with Incomplete Information Played by 'Bayesian' Players, I-III. Part I. The Basic Model." *Management Science* 14 (3): 159–82.

Hernandez-Leal, Pablo, Michael Kaisers, Tim Baarslag, and Enrique Munoz de Cote. 2019. "A Survey of Learning in Multiagent Environments: Dealing with Non-Stationarity." arXiv:1707.09183. Preprint, arXiv, March 11. https://doi.org/10.48550/arXiv.1707.09183.

Hirblinger, Andreas T. 2022. *When Mediators Need Machines (and Vice Versa): Towards a Research Agenda on Hybrid Peacemaking Intelligence*. January 25. https://doi.org/10.1163/15718069-bja10050.

Hong, Lu, and Scott E. Page. 2004. "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers." *Proceedings of the National Academy of Sciences of the United States of America* 101 (46): 16385–89. https://doi.org/10.1073/pnas.0403723101.

Horowitz, Michael C. 2018. "Artificial Intelligence, International Competition, and the Balance of Power." *Texas National Security Review*, May 15. https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/.

Ikenberry, G. John. 2009. *After Victory: Institutions, Strategic Restraint, and the Rebuilding of Order After Major Wars*. Princeton University Press.

Janis, Irving Lester. 1972. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton, Mifflin.

Jensen, Benjamin, Ian Reynolds, Yasir Atalan, et al. 2025. "Critical Foreign Policy Decisions (CFPD)-Benchmark: Measuring Diplomatic Preferences in Large Language Models." arXiv:2503.06263. Preprint, arXiv, March 8. https://doi.org/10.48550/arXiv.2503.06263.

Jervis, Robert. 1976. *Perception and Misperception in International Politics*. Princeton University Press.

Kaplan, Morton A. 1952. "An Introduction to the Strategy of Statecraft." *World Politics* 4 (4): 548–76. https://doi.org/10.2307/2008966.

Keeney, Ralph L., and Howard Raiffa. 1993. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Cambridge University Press.

Keohane, Robert O. 1988. "International Institutions: Two Approaches." In *International Organization*. Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315251981-7/international-institutions-two-approaches-robert-keohane.

Kersten, Gregory, and Hsiangchu Lai. 2007. "Negotiation Support and E-Negotiation Systems: An Overview." *Group Decision and Negotiation* 16 (October): 553–86. https://doi.org/10.1007/s10726-007-9095-5.

Koremenos, Barbara. 2001. "Loosening the Ties That Bind: A Learning Model of Agreement Flexibility." *International Organization* 55 (2): 289–325.

Koremenos, Barbara, Charles Lipson, and Duncan Snidal. 2001. "The Rational Design of International Institutions." *International Organization* 55 (4): 761–99.

Koremenos, Barbara, Charles Lipson, and Duncan Snidal. 2003. *The Rational Design of International Institutions*. Cambridge University Press.

Kydd, Andrew H. 2005. *Trust and Mistrust in International Relations*. Princeton University Press.

Lake, David A., and Robert Powell, eds. 1999. *Strategic Choice and International Relations*. Princeton University Press.

Lam, Remi, Alvaro Sanchez-Gonzalez, Matthew Willson, et al. 2023. "Learning Skillful Medium-Range Global Weather Forecasting." *Science* 382 (6677): 1416–21. https://doi.org/10.1126/science.adi2336.

Lamparth, Max, Anthony Corso, Jacob Ganz, Oriana Skylar Mastro, Jacquelyn Schneider, and Harold Trinkunas. 2024. "Human vs. Machine: Behavioral Differences between Expert Humans and Language Models in Wargame Simulations." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (1): 807–17. https://doi.org/10.1609/aies.v7i1.31681.

Lax, David A., and James K. Sebenius. 1986. "Interests: The Measure of Negotiation." *Negotiation Journal* 2 (1): 73–92. https://doi.org/10.1111/j.1571-9979.1986.tb00339.x.

Lax, David A., and James K. Sebenius. 2006. *3-d Negotiation: Powerful Tools to Change the Game in Your Most Important Deals*. Harvard Business Review Press.

Lin, Raz, Sarit Kraus, Tim Baarslag, Dmytro Tykhonov, Koen Hindriks, and Catholijn M. Jonker. 2014. "Genius: An Integrated Environment for Supporting the Design of Generic Automated Negotiators." *Computational Intelligence* 30 (1): 48–70. https://doi.org/10.1111/j.1467-8640.2012.00463.x.

Littman, Michael L. 1994. "Markov Games as a Framework for Multi-Agent Reinforcement Learning." In *Machine Learning Proceedings 1994*, edited by William W. Cohen and Haym Hirsh. Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-335-6.50027-1.

Liu, Nelson F., Kevin Lin, John Hewitt, et al. 2023. "Lost in the Middle: How Language Models Use Long Contexts." arXiv:2307.03172. Preprint, arXiv, November 20. https://doi.org/10.48550/arXiv.2307.03172.

Lukes, Steven. 2021. *Power: A Radical View*. Bloomsbury Academic.

Magee, Bryan, and Isaiah Berlin. 1978. *Men of Ideas: Some Creators of Contemporary Philosophy*. British Broadcasting Corporation.

Manor, Ilan. 2026. "From Digital Diplomacy to AI-Driven Diplomacies? Mapping the Potential Impact of AI on Diplomacy." *Communication and the Public*, April 2, 20570473261438571. https://doi.org/10.1177/20570473261438571.

March, James G., and Johan P. Olsen. 1983. "The New Institutionalism: Organizational Factors in Political Life." *American Political Science Review* 78 (3): 734–49. https://doi.org/10.2307/1961840.

Mell, Johnathan, and Jonathan Gratch. 2017. "Grumpy & Pinocchio: Answering Human-Agent Negotiation Questions through Realistic Agent Design." *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (Richland, SC), AAMAS '17, May 8, 401–9. https://dl.acm.org/doi/10.5555/3091125.3091186.

Mell, Johnathan, Jonathan Gratch, Tim Baarslag, Reyhan Aydoğan, and Catholijn M. Jonker. 2018. "Results of the First Annual Human-Agent League of the Automated Negotiating Agents Competition." *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (New York, NY, USA), IVA '18, November 5, 23–28. https://doi.org/10.1145/3267851.3267907.

Mellers, Barbara, Lyle Ungar, Jonathan Baron, et al. 2014. "Psychological Strategies for Winning a Geopolitical Forecasting Tournament." *Psychological Science* 25 (5): 1106–15. https://doi.org/10.1177/0956797614524255.

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, et al. 2022. “Human-Level Play in the Game of *Diplomacy* by Combining Language Models with Strategic Reasoning.” *Science* 378 (6624): 1067–74. https://doi.org/10.1126/science.ade9097.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, et al. 2015. “Human-Level Control Through Deep Reinforcement Learning.” *Nature* 518 (7540): 529–33. https://doi.org/10.1038/nature14236.

Nash, John F. 1950. “The Bargaining Problem.” *Econometrica* 18 (2): 155–62. https://doi.org/10.2307/1907266.

Ng, Andrew Y., and Stuart J. Russell. 2000. “Algorithms for Inverse Reinforcement Learning.” *Proceedings of the 17th International Conference on Machine Learning*, 663–70.

North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance*. Political Economy of Institutions and Decisions. Cambridge University Press. https://doi.org/10.1017/CBO9780511808678.

North, Douglass C. 1991. “Institutions.” *Journal of Economic Perspectives* 5 (1): 97–112. https://doi.org/10.1257/jep.5.1.97.

Noy, Shakked, and Whitney Zhang. 2023. “Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence.” *Science* 381 (6654): 187–92. https://doi.org/10.1126/science.adh2586.

Odell, John S. 2000. *Negotiating the World Economy*. Cornell University Press.

Parasuraman, Raja, and Dietrich H. Manzey. 2010. “Complacency and Bias in Human Use of Automation: An Attentional Integration.” *Human Factors* 52 (3): 381–410. https://doi.org/10.1177/0018720810376055.

Payne, Kenneth. 2026. “AI Arms and Influence: Frontier Models Exhibit Sophisticated Reasoning in Simulated Nuclear Crises.” arXiv:2602.14740. Version 1. Preprint, arXiv, February 16. https://doi.org/10.48550/arXiv.2602.14740.

Pouliot, Vincent, and Jérémie Cornut. 2015. “Practice Theory and the Study of Diplomacy: A Research Agenda.” *Cooperation and Conflict* 50 (3): 297–315. https://doi.org/10.1177/0010836715574913.

Pozniak, Charles, and Jeba Sania. 2026. “The Foreign Policy AI Evaluation Gap.” arXiv:2607.02955. Preprint, arXiv, July 3. https://doi.org/10.48550/arXiv.2607.02955.

Putnam, Robert D. 1988. “Diplomacy and Domestic Politics: The Logic of Two-Level Games.” *International Organization* 42 (3): 427–60.

Raiffa, Howard, John Richardson, and David Metcalfe. 2002. *Negotiation Analysis: The Science and Art of Collaborative Decision Making*. Harvard University Press.

Ringquist, John. 2025. “Using CamoGPT AI to Build Scenario-Driven Change: The Example of Guam as a Hybrid Attack Target.” Army University Press, Military Review, July. https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/July-August-2025/CamoGPT-AI/.

Rivera, Juan-Pablo, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. 2024. “Escalation Risks from Language Models in Military and Diplomatic Decision-Making.” *The 2024 ACM Conference on Fairness Accountability and Transparency*, June 3, 836–98. https://doi.org/10.1145/3630106.3658942.

Samore, Gary. 2015. *The Iran Nuclear Deal: A Definitive Guide*. Belfer Center for Science and International Affairs, Harvard Kennedy School. https://www.belfercenter.org/publication/iran-nuclear-deal-definitive-guide.

Sartori, Anne E. 2005. *Deterrence by Diplomacy*. Princeton University Press.

Satter, Raphael, Humeyra Pamuk, Raphael Satter, and Humeyra Pamuk. 2025. “US State Department Cable Says Agency Using AI to Help Staff Job Panels.” United States. *Reuters*, June 9. https://www.reuters.com/world/us/us-state-department-cable-says-agency-using-ai-help-staff-job-panels-2025-06-09/.

Schelling, Thomas C. 1960. *The Strategy of Conflict: With a New Preface by the Author*. Harvard University Press.

Shoham, Yoav, Rob Powers, and Trond Grenager. 2007. "If Multi-Agent Learning Is the Answer, What Is the Question?" *Artificial Intelligence*, Foundations of Multi-Agent Learning, vol. 171 (7): 365–77. https://doi.org/10.1016/j.artint.2006.02.006.

Silver, David, Aja Huang, Chris J. Maddison, et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (7587): 484–89. https://doi.org/10.1038/nature16961.

Silver, David, Thomas Hubert, Julian Schrittwieser, et al. 2018. "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play." *Science* 362 (6419): 1140–44. https://doi.org/10.1126/science.aar6404.

Simon, Herbert A. 1962. "The Architecture of Complexity." *Proceedings of the American Philosophical Society* 106 (6): 467–82.

Sullivan, Jake, and Tal Feldman. 2026. "Geopolitics in the Age of Artificial Intelligence." *Foreign Affairs*, January 27. https://www.foreignaffairs.com/united-states/geopolitics-age-artificial-intelligence.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning, Second Edition: An Introduction*. Bradford Books.

Tessler, Michael Henry, Michiel A. Bakker, Daniel Jarrett, et al. 2024. "AI Can Help Humans Find Common Ground in Democratic Deliberation." *Science* 386 (6719): eadq2852. https://doi.org/10.1126/science.adq2852.

Tetlock, Philip E., and Dan Gardner. 2015. *Superforecasting: The Art and Science of Prediction*. Superforecasting: The Art and Science of Prediction. Crown Publishers/Random House.

Thompson, James D. 1967. *Organizations in Action: Social Science Bases of Administrative Theory*. Routledge.

Toft, Monica Duffy. 2006. "Issue Indivisibility and Time Horizons as Rationalist Explanations for War." *Security Studies* 15 (1): 34–69. https://doi.org/10.1080/09636410600666246.

Tversky, Amos, and Daniel Kahneman. 1973. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5 (2): 207–32. https://doi.org/10.1016/0010-0285(73)90033-9.

Vassilev, Apostol, Alina Oprea, Alie Fordyce, Hyrum Anderson, Xander Davies, and Maia Hamin. 2025. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. NIST Artificial Intelligence (AI) 100-2 E2025. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-2e2025.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Vinyals, Oriol, Igor Babuschkin, Wojciech M. Czarnecki, et al. 2019. "Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning." *Nature* 575 (7782): 350–54. https://doi.org/10.1038/s41586-019-1724-z.

Von Neumann, John, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton University Press.

Waltz, Kenneth N. 2010. *Theory of International Politics*. Waveland Press.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, et al. 2023. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." arXiv:2201.11903. Preprint, arXiv, January 10. https://doi.org/10.48550/arXiv.2201.11903.

Welch, Ivo. 2000. “Herding among Security Analysts.” *Journal of Financial Economics* 58 (3): 369–96. https://doi.org/10.1016/S0304-405X(00)00076-3.

Zartman, I. William, and Maureen R. Berman. 1982. *The Practical Negotiator*. Yale University Press.

Zhang, Xiangyuan, Kaiqing Zhang, Erik Miehling, and Tamer Basar. 2019. *Non-Cooperative Inverse Reinforcement Learning*. September 6. https://openreview.net/forum?id=SJep04rlIH.