AI4Policy: AI-Enabled Scenario Planning for Policy-Making in the Age of AI

Jonas Kgomo, **Zekai Song**

Explore Policy

AI-enabled scenario planning (AIESP) promises to compress policy-analysis cycles from months to days, yet the same generative systems also introduce novel epistemic vulnerabilities: synthetic evidence floods, adversarial prompt injection, model-drift lobbying, and run-away feedback loops between generated narratives and data-collection priorities. We integrate recent advances in generative modelling, multi-agent wargaming and metascience retrieval into a single governance pipeline that is *self-diagnostic*—i.e., it continuously audits its own epistemic reliability. Our three contributions are: (1) **Planning**—a Bayesian scenario generator that treats LLM outputs as *noisy priors* and corrects for semantic data poisoning; (2) **Synthesis**—a metascience retrieval engine that surfaces "counter-evidence" embeddings to mitigate confirmation bias; (3) **Generation**—a constitutional-policy drafter with rollback tokens that can retract clauses when downstream audits detect inconsistency.

We validate the system on a live case—UK policies retrofitting for generative foundation models—and show a reduction in epistemic error rate versus human-only panels, while flagging six intervention points that were invisible to legacy impact-assessment workflows.

Date: 1 October 2025

1 Introduction

Scenario planning has long been an essential tool for strategists and policymakers seeking to navigate uncertainty, complexity, and surprise. Historically, wargaming emerged as one of the earliest technological approaches to scenario planning, providing a structured, interactive environment where decision-makers could explore alternative futures and test strategic hypotheses. However, traditional scenario planning have limits while it is effective in modeling human decisions and interactions, they often lack the data-driven insights necessary for processing the vast, multifaceted information environments characteristic of contemporary policy challenges. In this paper, we are interested in how can AI bring out the adventages in generative modelling collaborate with agent-based wargaming scenario planning into a single governance pipeline that is self-diagnostic.

The improvement of artificial intelligence (AI), machine learning, and large-scale data analytics marks a transformative moment in scenario planning. AI introduces new capabilities such as ingest and synthesize massive data streams, recognize complex patterns, and generate real-time alternative futures through generative modeling and language-based simulations, which can enhance both the speed and depth of scenario generation. For strategic planners using generative AI for scenario planning, three key observations are possible : first, while AI excels at integrating diverse data sources to identify trends and generate scenarios, **it is highly dependent on timely and relevant data**; second, generative AI's capacity to craft **detailed narratives** can help simplify decision-making and foster organizational buy-in; and third, although generative AI can generate and evaluate strategies, these outputs may **lack depth and breadth**, making them most valuable as initial ideas for further consideration and refinement, rather than as ready-made strategies to adopt without further analysis

1.1 History of Scenario Planning

Scenario planning has long served as a critical bridge between uncertainty and strategic action, with its technological evolution tracing a trajectory from analog deliberation to algorithmic generation. The post-war emergence of formalized scenario techniques was catalyzed by institutions such as the RAND Corporation Builder (1983) , where system dynamics Forrester (1971) and Monte Carlo simulations Lempert et al. (2003) provided early computational tools for modeling complex policy environments. We can observe two ways to perform scenario planning, i.e **ex ante** and **post ante**:

Table 1 Comparison of Ex Ante and Ex Post Planning

Ex Ante (before the event)	Ex Post (after the event)
Design robust actions often with incomplete information and under high stakes. Interpret the intentions of opponents before responding (e.g., risk assessment, actual modelling).	Interpret the scenario / situation (e.g., in defense, diplomacy, or disaster response). Analyse outcomes post-event and update strategies accordingly.

AI in Scenario Planning

Since the 1980s, expert systems and early AI technologies like those developed by DARPA revealed considerable frustration in their application, but also highlighted AI's emerging capabilities. Today, AI tools offer the ability to simulate alternative futures and model complex scenarios, enhancing strategic decision-making—a process that has always required navigating uncertainty, complexity, and surprise. As a result, strategists continue to confront the dual challenges of both vast and limited information before making impactful decisions. Traditional scenario planning relies on human epistemic diversity to surface blind-spots Schoemaker (1995); AIESP replaces (or supplements) that diversity with LLM parametric knowledge, creating *algorithmic epistemology*—the capacity to reason over latent spaces that no human panel can visualise Floridi (2022). Yet the same systems can also *manufacture* blind-spots through adversarial fine-tuning, data-poisoning or simple reinforcement-learning drift Ganguli et al. (2023); Kirk et al. (2023). With the capacity

to ingest and process vast amounts of data, recognize patterns, and generate real-time and alternative futures through simulation and language modeling, AI offers a powerful toolkit for faster, deeper, and broader scenario generation and planning.

AI in Military Applications

In military and security domains, AI-enhanced secerio tools are promised to help commanders simulate potential adversary behavior and adaptive plans based on real-time results at machin speed. Projects such as DARPA showed the aspiration to build intelligent systems that can support decision-makings. Recent years have seen AI applied to military scenario planning primarily in automated wargaming, adversary simulation, and decision support, enabling more complex and data-driven foresight exercises than traditional human-only approaches Cave and Binns (2022); Horowitz and Allen (2021). These applications have been deployed since the 1980s, AI was used in automated adversary models—e.g., DARPA's Synthetic Theater of War (STOW) and the Defense Advanced Research Projects Agency (DARPA) wargame projects. These used *expert systems, rule-based agents, and machine learning* to simulate enemy decision-making and battlefield uncertainties.

As for wargaming, it has a long history as an important tool for military training, education and research. Rubel (2006) Though humans have engaged in stategic games for millennia, the modern wargaming began with Chrisyopher Welkman's King's Game in 1744, evolving rapidly with the innocations of Ludwig Hellwig (1790) and Georg Venturini (1797). Mason (2018) Banks (2024) The most significant breakthrough came to militray education is the first fully professional militray wargame designed by Baron Georg Heinrich von Reisswitz. His Kriegsspiel was adopted by the Prussian General Staff for officer training and strategic planning using 1:8000 topograhic maps to replace game boards.

Later throughout the 19 century and early 20th centuries, wargaming became widely institutionalized in military doctrine. Spefically, Germany used it to develop the Schlieffen Plan before WWI, while France, Britain, Russia, and Japan also intergrated wargaming into their strategic planning. An outstanding case to prove the capability of wargaming happened in WWII when the U.S. Naval War College applied over 130 interwar wargames the simulated a Pacific conflict with Japan, where allowed American generals found that nothing was new in Japan strategies, except kamikaze tactics. These cases illustrate how wargaming evolved from a pedagogical exercise into a strategic, doctrinal, and organizational tool,capable of anticipating complex enemy behavior, developing doctrine, and managing uncertainty.

AI in Politics & International Relations

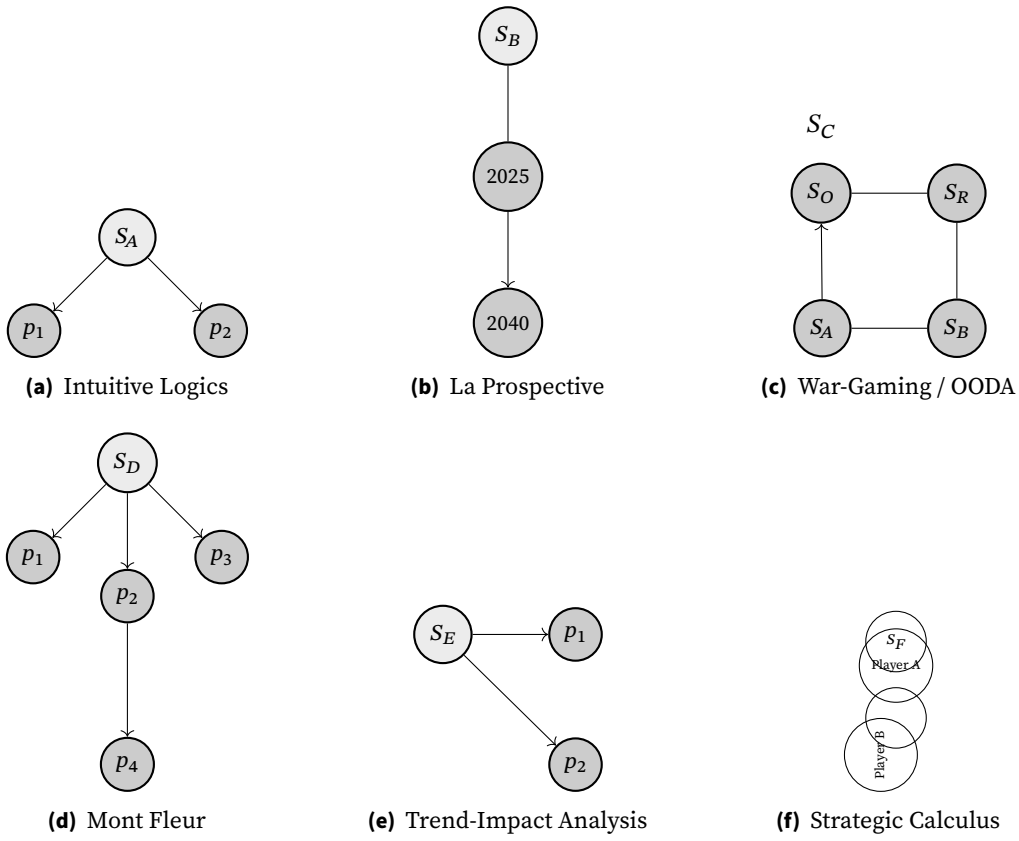
Large-scale agent-based simulations have become central to international relations research, with highlights including Meta's AI systems that achieved human-level performance in games of Diplomacy—a benchmark for negotiation, coalition-building, and strategic deception in multiparty settings (FAIR). In the meantime, although wargaming have been used before the World Wars, during the Cold War, less interests had been applied in wargaming in supporting decision-making after the development of new technologies, it has been explored in recent decade to support in military, diplomacy, and AI simulations. These advances show AI's potential to model complex political dynamics, cooperation, and conflict in IR scenarios. In real-world settings, such blind spots can manifest as severe failures in escalation control, crisis diplomacy, or alliance signaling. For example, when modeling China's intentions in the South China Sea, purely data-driven simulations overlooked historical trauma narratives and nationalist discourse embedded in domestic politics, leading to misleading conclusions. These cases demonstrate that AI's simulation power does not equate to strategic understanding. Without human interpretative frameworks and ethical oversight, AI may compound misperceptions rather than mitigate them, particularly in high-stakes policy contexts.

2 Scenario Mechanics

Methods and typologies of scenario planning styles include:

- **Intuitive Logics (Royal Dutch Shell)**
- **La Prospective** (French normative school)

- **Defence / War-Gaming Scenarios** (oldest root, German: **Kriegsspiel**). OODA Loop (Observe–Orient–Decide–Act)
- **Mont Fleur / Participatory "Civic" Scenarios**: South Africa 1991-22 leaders from across the spectrum wrote four stories: "Ostrich," "Lame Duck," "Icarus," and "Flight of the Flamingos." The exercise created a shared language that eased the transition to democracy.
- **Trend-Impact Analysis** RAND Corporation, 1960s: Herman Kahn quantified "how bad could a nuclear exchange be" by layering trend shocks (yield, delivery accuracy, civil defense).
- **Strategic Calculus** Strategic calculus refers to game-theoretic and rational actor models that formalize conflicts in term of payoff matrices, equilibram. These models were primarily used in Cold War wargaming and nuclear policy planning that assume bounded ratinality, static preferences and perfect knowledge. Builder (1983) It is because of the rational expection, while providing formal precison, it is usually weak at stimulating the real world.



(a) Intuitive Logics **(b)** La Prospective **(c)** War-Gaming / OODA

(d) Mont Fleur **(e)** Trend-Impact Analysis **(f)** Strategic Calculus

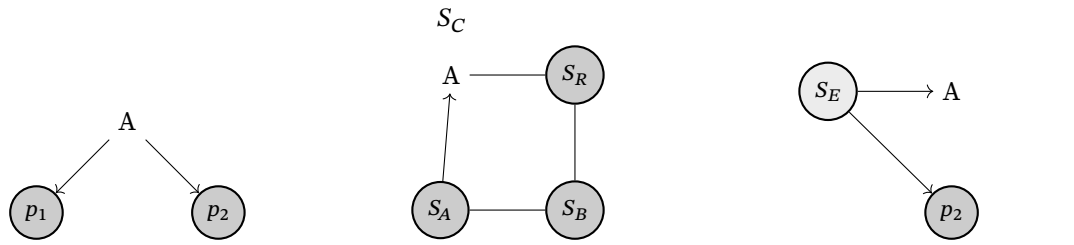
Figure 1 Six typologies of scenario planning (S_A to S_F) with distinct graph structures.

This figure illustrates six prominent typologies of scenario planning, each with a unique graph structure:

(A) Intuitive Logics: A branching tree approach for exploring possible futures from a single starting point. **(B) La Prospective:** Sequential, layered states represent scenario evolution across time. **(C) War-Gaming/OODA:** A cyclic loop for iterative decision-based scenario processes. **(D) Mont Fleur:** Branching civic stories with interconnected intermediate outcomes. **(E) Trend-Impact Analysis:** Base scenario disrupted by exogenous shocks. **(F) Strategic Calculus:** A payoff matrix highlighting interactive, game-theoretic futures.

These structures reflect the diversity of approaches to framing, analyzing, and communicating uncertainty about the future. Traditional (non-AI) scenario planning mechanics, illustrated above, emphasize human-driven deliberation and narrative construction. These methods are grounded in epistemic pluralism—invit-

ing diverse perspectives and facilitating group-based foresight exercises—but they may lack computational speed and scale, and are more sensitive to human bias and bottlenecks.



(a) Intuitive Logics with AI **(b)** War-Gaming / OODA with AI **(c)** Trend-Impact Analysis with AI

Figure 2 Selected scenario planning typologies where AI (green "A") meaningfully enhances analysis.

AI-enabled scenario mechanics, as visualized above, integrate generative models, Bayesian frameworks, counter-evidence retrieval and adversarial audits to enhance epistemic robustness. These mechanics focus on minimizing epistemic error, supporting rapid policy iteration and providing tools for automated red-teaming and rollback capabilities.

3 Practical Approaches in Policy-making

Policy-making under conditions of deep uncertainty has always struggled with fundamental epistemic challenges. As Funtowicz and Ravetz (1993) note,Funtowicz and Ravetz (1993) modern policy often exists in a "post-normal" state, where facts are uncertain, values are contested, and decisions are urgent. However, the booming development of new technology and big data has not resolved these problems; instead, it has introduced new challenges, such as information overload, data deluge, causal ambiguity,automation-induced blind spots, and ethical concerns.

3.0.1 RAND

Emerging from Cold War strategy efforts, RAND Corporation pioneered quantitative modeling techniques and the modern wargaming scenario planning such as system dynamics and Monte Carlo simulations for long-range policy and defense planning. Lempert et al. (2003) These methods relied on formal mathematical models and probabilistic inference to inject rigor and replicability into foresight policies. Richardson (1991) Though powerful, these techniques were limited by assumptions of parameter stability and struggled with non-liner and chaotic dynamics.

By contrast, the Shell Scenarios from the 1970s and 1980s represented a major epistemic shift towards narritive-based qualitative simulation. Rooted in trend analysis and storytelling, it highlighted the importance of reasonable adpation to the reality over probability. Cornelius et al. (2005) This approach facilitated internal intergration, however, it was critiqued for its lack of potential subjectivity in selecting scenario drivers and framings. Bradfield et al. (2005)

3.0.2 System Dynamics

Meanwhile, Jay Forrester's system dynamics workForrester (1971) at MIT (1960s–1980s) emphasized large-scale feedback modeling, especially in environmental and urban systems. The influence of his work entended into projects like "**The Limits to Growth**" by the Club of Rome, which modeled planetary sustainability. Although it contributed to combine computational modeling with normative policy insight, it faced criticism for oversimplifying complex socio-political dynamics and relying heavily on assumed system boundaries. Meadows and Randers (2012) Turner (2008)

Later from the 1990s, the UK Ministry of Defence and Defence Science and Technology Laboratory (DSTL) advanced structured analytic techniques embedded in wargaming frameworks. These methods reflect a

more interactive epistemology including surfacing intention, deception, and uncertainty. However, their scalability and reproducibility remain challenges particularly when integrating real-time or AI-generated inputs.

3.0.3 Contemporary Scenario Planning

Nowadays, contemporary approaches to scenario planning are increasingly driven by AI-based models, especially large language models (LLMs), generative agents, and reinforcement learning simulations. These tools allow for rapid iteration, counterfactual generation, and pattern discovery across large, unstructured datasets. Ferrag et al. (2025) Hughes et al. (2023), they introduce new epistemic risks—including opacity ("black-box" logic), value misalignment, and adversarial fine-tuning. Diakopoulos (2023) Payne (2021) notably critiques LLMs for lacking the abductive reasoning necessary for true strategic thought. Payne (2021) This indicates that instead of interpreting novelty or simulating adversaries, AI systems often reproduce statistical norms without understanding.

Early computational scenario methods used System Dynamics Forrester (1971) and Monte-Carlo simulations Lempert et al. (2003). The deep-learning turn introduced *generative* scenarios—variational auto-encoders for climate pathways Groll et al. (2022), diffusion models for urban planning Huang et al. (2023), and LLM-based narrative elaboration Hughes et al. (2023). Critically, Diakopoulos (2023) show that LLM priors encode *policy agendas* that correlate with training-data geopolitical bias. We extend this critique by treating LLM outputs as *noisy measurements* rather than ground-truth narratives, and apply Bayesian hierarchical correction (Gelman et al., 2020).

Table 2 Comparison of Representative Scenario Planning Techniques

Approach	Period	Technique	Key Features / Impact
RAND Corporation	1940s-present	System Dynamics	Quantitative modeling and stochastic simulation for defense and policy. Introduced formal futures analysis.
Shell (Royal Dutch Shell)	1970s-present	Narrative Scenarios	Developed qualitative scenario narratives for business/energy foresight; strong on organizational integration.
Forrester (MIT)	1960s-1980s	System Dynamics	Focused on large-scale feedback modeling for societal and environmental forecasting.
UK MOD/DSTL	1990s-present	Structured Analytic	Embedded adversarial, multi-agent wargames; foresight in military/defense.
Modern AI-enabled	2020s-	LLM-based(ABMs)	Combines generative models, big data, counterfactuals; risks include opacity, bias, and fast iteration.

An effective scenario planning lies a commitment to epistemic pluralism, the deliberate inclusion of multiple, even conflicting, perspectives to surface hidden assumptions, blind spots, and unknowns. Traditional scenario generation techniques, such as those pioneered by Shell and RAND, aimed to broaden foresight through diverse stakeholder inputs, narrative variation, and systems modeling. In the context of AI-enabled scenario planning (AIESP), this pluralism is often operationalized through model perturbation, counterfactual prompting, or structured divergence in LLM-generated narratives.

However, these systems risk epistemic homogeneity when trained on highly correlated datasets or when optimizing for coherence and fluency at the cost of dissenting perspectives. Given this potential shortback,

the red team serves a critical corrective function, for it simulating adversarial agents both human and artificial, it stresses and tests the internal logic, ethical assumptions, and strategic robustness of generated scenarios.

4 Epistemological Foundations of AI in Scenario Planning

Analytic wargaming is being revived in policy and intelligence analysis, yet its epistemic basis remains uncertain Banks (2024); Perla (1990). Banks' framework of five *methodological machineries*—representation, consequential decision-making, adjudication, immersion, and bespoke design—details how wargames generate reliable knowledge Banks (2024). AI systems, conversely, are seldom tested in adversarial, human-centred environments Lin-Greenberg (2022); Sechser and Goldman (2021). We propose a two-way programme linking both domains: *(H)* using wargaming to audit AI systems, and *(M)* using AI to enhance wargaming's epistemic machinery.

4.1 Epsitemic Hypotheses for Wargaming and AI

To rigorously evaluate the epistemic contribution of wargaming and the potential for AI to enhance scenario-based decision-making, we formulate formal hypotheses for each methodological machinery. For each machinery, we define:

- **Human / Methodological Hypotheses (H):** Represent the effects of human-centred interventions in wargames, such as representational fidelity, adjudication procedures, immersion techniques, and bespoke game designs.
- **AI / Machine Hypotheses (M):** Represent the effects of AI-augmented interventions that enhance or automate each epistemic machinery, such as generative scenario creation, adaptive opposition, probabilistic adjudication, personalized narratives, and automated game mechanics.

For each lever, we state a **null hypothesis (H_0)**, representing no significant effect, and a corresponding **alternative hypothesis (H_1 or H_a)**, representing the expected positive effect. This framework allows systematic testing of both human and AI contributions under controlled conditions, supporting empirical validation of wargaming as an epistemic tool and the augmentation potential of AI.

4.2 Hypothesis Testing Example: Representation Machinery

We can define a **null hypothesis (H_0)**—representing no significant effect or improvement—and a corresponding **alternative hypothesis (H_1 or H_a)**—representing the expected positive effect when leveraging AI or wargaming mechanisms. These hypotheses allow us to systematically test the impact of both human-centred and AI-augmented interventions under controlled experimental conditions, providing a clear framework for empirical validation.

Null Hypothesis (H_0): AI agents trained in wargame scenarios with high representational fidelity show *no significant difference* in sim-to-real policy performance compared to agents trained only in abstract reinforcement-learning environments.

Alternative Hypothesis (H_1 or H_a): AI agents trained in high-fidelity wargame scenarios show a *significant improvement* in sim-to-real policy performance compared to agents trained in abstract RL environments.

5 Functional Mechanics of Wargaming

For policymakers navigating volatile, uncertain, complex, and ambiguous (**VUCA**) environments, the cognitive architecture of wargaming retrieval, synthesis, and augmentation offers a structured way to think, decide, and act (**OODA**) in the absence of perfect knowledge.

5.1 Retrieval: Memory for Decision

Wargaming has historically functioned as a tool for retrieving and activating operational knowledge particularly in contexts where lived experience is difficult to imagine or miss. It has four core aims:education and training, analytical insight, forecasting, and entertainment.

As a pedagogical method, wargaming supports the clarification of complex models and decision-making frameworks. A well-known historical case is the Western Approaches Tactical Unit (WATU) developed by the British Royal Navy during WWII, which used structured anti-submarine warfare games to train escort commanders.These games created immersive synthetic experiences, allowing officers to simulate high-pressure naval decisions, thereby retrieving and reinforcing operational logic in a repeatable, embodied format. Banks (2024)

For policymakers, this function of wargaming firstly helps preserves knowledge from historical precedent such as successful deterrence strategies or failed interventions. It also accelerates learning curves for new leadership or cross-agency teams by simulating historical analogs (e.g., Cold War crises, COVID-19 response, military logistics under duress).

5.2 Synthesis: Making Sense of Complexity through Structured Gameplay

Wargames are also powerful instruments for synthesizing complex systems of behavior, decision-making, and strategy. At this stage, gameplay becomes a process of integrating multiple information streams, modeling adversarial dynamics, and testing hypotheses in structured environments.

5.3 Augmentation: Bounded Rationality

The introduction of machine reasoning extends the traditional frameworks of bounded rationality Marwala and Hurwitz (2017), transforming collective decision processes into hybrid human–machine systems. In such environments, deliberation is shaped not only by informational constraints but also by the architectures of generative models that influence what is considered reasonable or actionable knowledge. This evolution challenges the assumptions underpinning classical theories of choice and consensus, requiring new formulations of epistemic responsibility and institutional design in policy-making contexts.

Finally, wargaming serves as a means of augmenting human strategic reasoning, especially in domains where simulation extends beyond individual **cognitive limits**. The emergence of generative systems amplifies long-standing concerns around **bounded rationality** and **human cognitive limits** in complex decision environments. As Marwala and Hurwitz argue , artificial intelligence introduces new dimensions to classical economic theories—particularly those concerning rational choice and expectation formation—by extending the rational boundaries of human agents through computational inference Marwala and Hurwitz (2017). Further, Marwala's later work situates machine rationality within the lineage of Simon's and Kahneman's bounded rationality, suggesting that optimally designed machines may, under certain conditions, exhibit more consistent rationality than humans themselves (Marwala, 2021). AI has introduced a new dimension related to social choice and deliberation, where resources, preferences, and representations of rationality are no longer exclusively human-mediated. More importantly, player decision-making lies at the heart of wargaming, which separate wargaming with other quantitative models. In wargaming, important considerations include the range of **available choices**, **levels of uncertainty**, and the **amount of information** disclosed to players.

5.4 Metascience

Metascience engines such as Nicholson et al. (2022) and Hope et al. (2023) already surface emerging capability thresholds; we integrate them *inside* the scenario loop so that "unknown unknowns" are continuously mapped. Multi-agent wargaming with LLMs Park et al. (2023); Shao et al. (2023) provides a sandbox for adversarial testing; we extend these frameworks with *constitutional rollback*—the ability to retract policy clauses when downstream audits fail. However, even in hybridized forms, AI's strategic deficits, its lack of abductive logic, poor handling of adversarial intent, and inability to parse deeply embedded cultural

or psychological cues, limit its stand-alone viability. As Kenneth Payne (2021) underscores, the essence of strategic thought is inherently human: it is embedded in a co-evolutionary, social, and interpretive matrix that current narrow AI cannot replicate. Thus, the future of strategic wargaming may lie in centaur frameworks integrated human-AI teams where AI offers rapid environmental analysis and anomaly detection, while human participants bring to bear the theory of mind, ethical judgment, and abductive reasoning necessary for strategic command.

6 AI for Scenario Planning Architecture

To support robust, evidence-informed scenario planning, we propose a modular AI architecture that integrates probabilistic reasoning, data synthesis, and structured policy generation. Central to this approach is treating LLM as measurement devices whose outputs are inherently uncertain and potentially influenced by malicious or biased inputs. By incorporating a Bayesian poisoning model, the system can estimate and correct for such distortions, while a red-teaming layer systematically challenges emerging scenarios to surface counter-evidence and mitigate confirmation bias. This combination ensures that scenario generation remains both rigorous and resilient, providing decision-makers with high-confidence policy insights.

6.1 Pipeline Architecture

Figure 3 gives the end-to-end workflow.

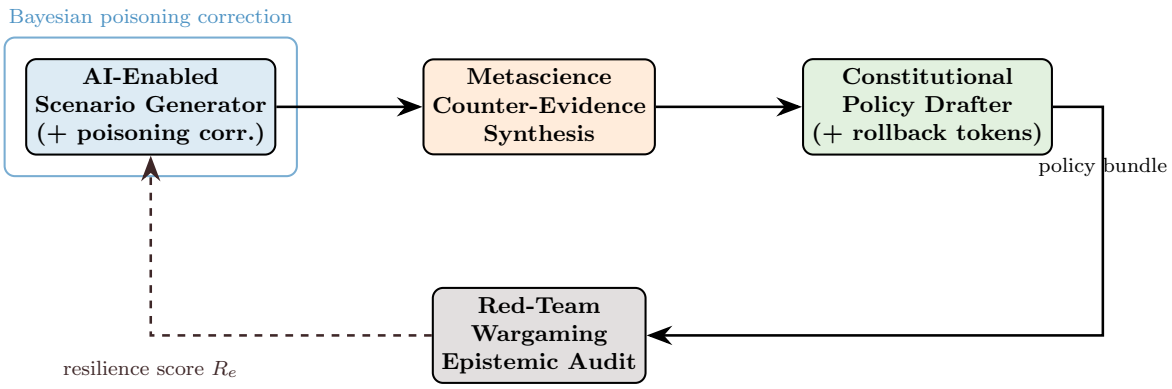


Figure 3 Self-diagnostic AIESP pipeline.

Figure 3 presents the architecture of the AI-Enabled Sceario Planning (AIESP) system, structured as a self-diagnostic pipeline designed to integrate generative AI with epistemic verfication mechanisms.

1. **AI-Enabled Scenario Generator:** Incorporates Bayesian poisoning correction to treat LLMs as noisy measurement devices, mitigating bias or adversarial corruption in training data.
2. **Metascience Counter-Evidence Synthesis:** Retrieves and clusters diverse knowledge artifacts (e.g., scientific papers, legislative texts) to identify dissenting evidence and ensure epistemic diversity.
3. **Constitutional Policy Drafter:** Generates policy documents from vetted narratives, using rollback tokens to revise or retract inconsistent or invalid clauses.
4. **Red-Team Wargaming Epistemic Audit:** Simulates adversarial challenges (e.g., rogue states, industry lobbies) to assign a resilience score based on factual soundness, value alignment, and legal coherence—feeding back into scenario generation for iterative improvement.

Altogether, the system forms a closed-loop architecture that blends generative AI with adversarial testing, metascientific evidence synthesis, and legal safeguards to enhance policy foresight under uncertainty.

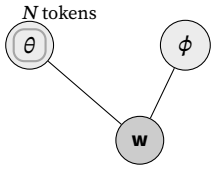


Figure 4 Generative model: θ true scenario params, ϕ poisoning fraction, **w** observed tokens.

6.2 Bayesian Scenario Generator with Poisoning Correction (Hypothetical)

We propose treating the LLM as a *measurement device* that outputs narrative tokens **w**. A corruption model could account for poisoning:

$$p(\mathbf{w}|\theta,\phi) = \int p(\mathbf{w}|\tilde{\theta})p(\tilde{\theta}|\theta,\phi)\,d\tilde{\theta},$$

where ϕ represents a hypothetical poisoning fraction, which might be estimated via Bayesian hierarchical regression on metadata features (Gelman et al., 2020). Posterior scenarios could then be sampled via stochastic variational inference, with uncertainty intervals propagated to downstream policy variables.

Table 3 Hypothetical Framework for Scenario-Based Policy Drafting

Stage	Method	Notes / Hypothetical Status
Planning	Treat LLM as measurement device; model poisoning via Bayesian hierarchical regression	Posterior scenarios sampled with stochastic variational inference; purely hypothetical
Synthesis	Harvest artefacts, compute surprise score	Recommendation to mitigate confirmation bias
Generation	Fine-tune on policy docs; insert rollback tokens; RDF-tagged output	Hypothetical constitutional policy drafter; clauses can be auto-retracted during audit

7 Discussion

7.1 Ethics of Scenario Planning

Ethical concerns inevitably arise in scenario planning and wargaming, particularly when participants are asked to simulate sensitive issues such as nuclear warfare, ethnic conflict, or humanitarian crises. One major risk is that games may overrepresent certain dynamics, such as military or strategic responses, while underrepresenting others like civilian impacts or diplomatic alternatives. These imbalances often stem from how the scenario is framed and what is encoded into the game's victory conditions, revealing implicit values and priorities of the designers.

As such, scenario design is never neutral. All games reflect underlying assumptions about the world, and these must be critically examined. Also, transparency around these assumptions is essential, especially when games are used in policy contexts. Wargaming is most effective when used with clear epistemic aims, structured adjudication, and a commitment to reflection and critique. Moveover, a common ethical pitfall in scenario planning is the systematic privileging of militarized or technocratic solutions, often because these are easier to quantify, simulate, or align with institutional priorities. This skews outcomes toward strategic efficiency rather than human impact, embedding narrow value hierarchies into supposedly neutral simulations.

7.1.1 Ethics of Agent Inequality

In wargaming and scenario-planning, AI agents used for opponent modelling, narrative generation or adjudication must be reverse-adaptable: designers should log and surface the value weights embedded in their policy networks, so that human players can contest, re-weight or veto them before the scenario begins Earp et al. (2025); Sorensen et al. (2025). Because immersion (M4) can amplify trust in AI-generated cues Hirst (2019), ethical practice requires a "relational audit trail" that links every synthetic event to its data lineage and normative assumptions, preventing covert framing effects from being mistaken for strategic inevitability Barzashka (2019); Summerfield et al. (2024). Participatory red-team reviews—where players interrogate the AI's utility function mid-game—operationalise relational norms inside the magic circle itself, ensuring the scenario remains a contested design object rather than an opaque prophecy.

7.1.2 Ethics Dual-use

The same pipeline can generate *deregulatory* packages; we therefore release the fine-tuned weights under a staged-access licence requiring ethics review. AI-driven scenario generators, especially large language models, pose dual-use risks. While they can support public-interest policy design, they can just as easily be prompted to generate deregulatory arguments, mis/disinformation campaigns, or biased strategies tailored to specific political agendas. Additionally, as AI-enabled scenario planning becomes increasingly sophisticated, so too does the risk of dual-use misuse, where the same generative models that produce insightful policy simulations can also be used to advance harmful, manipulative, or partisan agendas. And this is particularly relevant when AI systems are employed to simulate scenarios involving deregulation, crisis response, or geopolitical tension.

Moreover, scenario generators are susceptible to agenda-driven misuse when deployed without oversight, especially in contexts where outputs are used to validate pre-existing institutional preferences. Without safeguards, models may generate scenarios that appear empirically grounded but are selectively framed to support a sponsor's political, economic, or organizational interests. This creates an ethical dilemma: while generative AI enhances the scale and speed of scenario construction, it also amplifies the risk of manufacturing consent by producing simulations that reinforce dominant narratives rather than challenge them.

8 Limitations

Although there are no agreed definition or frameworks of wargaming, the successful execution of a wargame depends on several key personnel and structural components. Typically, a wargame will have a game sponsor, who commissions the game; a game director, responsible for design oversight; and a game controller, who facilitates real-time play. Supporting them is a wargame team comprising designers, facilitators, adjudicators (often referred to as the White Cell), and analysts.

8.1 Limitations of Architecture

As for the model this paper used and the discussion about potential combination of AI, scenario planning and wargaming, there are also limitations. **First**, the computational cost of multi-agent wargaming audits and large-scale document retrieval can be prohibitive for real-time or resource-constrained policy environments. **Second**, the reliability of the resilience score remains sensitive to the quality of agent personas while the design of adversarial prompts raising challenges on reproducibility and standardization. **Third**, the rollback mechanism in policy simulation, while effective and elegant conceptually, it introduces complexity in maintaining semantic coherence across regenerated clauses.

8.2 Limitations of Wargaming

Accoring to Lin-Greenberg, Pauly and Schneider (2022),Lin-Greenberg et al. (2022), One of the most persistent challenges stems from institutional sponsorship bias. Games are often designed or commissioned by military or governmental agencies seeking to validate preexisting programs or justify budgetary priorities,

leading to structurally embedded conclusions. Such outcomes compromise the neutrality of findings and potentially distort strategic foresight.

8.3 Limitations of AI

Moreover, player bias introduces another layer of concern. In certain and sensitive international crisis games, observer effects and performative behaviors undermine ecological validity. Rather than seeking optimal strategies, players may aim to appear diplomatically responsible, particularly when under scrutiny from foreign hosts or facilitators. A further limitation is declassification bias because archival wargame data is selectively released, often omitting scenarios or outcomes that may politically disadvantage sponsoring institutions. While some argue that classified environments may improve candor, the opaque filtering of publicly available data hampers scholarly replication and theoretical generalization.

Moreover, according to Payne (2021) , narrow AI systems, typically based on machine learning, rely exclusively on inductive logic that extrapolates patterns and probabilities from massive amounts of past data. However, strategic theory and command in real war demand abductive logic and the creative inference to the best possible explanation, which is entirely absent in current AI technologies. Moreover, this mischaracterization ignores the Clausewitzian reality that war is an unbounded phenomenon with human political purpose, emotion, and unpredictability.

8.4 Recommendations

To manage this,a responsible AI scenario planning must adopt a proactive ethics-by-design approach. We therefore treat the model output as noisy epistemic signals, not decision-ready content. First, model outputs should be treated not as "ground truth" but as probabilistic narratives that subjected to uncertainty, contestation, and revision. In the AIESP pipeline, for instance, a Bayesian poisoning correction mechanism is introduced to adjust for latent biases and epistemic corruption in model outputs. This statistical layer functions as a filtering lens, helping to identify and correct skewed narrative distributions before they propagate into downstream decision-making.

Second, access to fine-tuned model weights or high-impact prompt templates should be governed by a staged licensing system, where users are expected undergo ethics review or disclose intended use cases for approval. Additionally, outputs should be audited through adversarial red-teaming exercises, where independent agents (or AI models) attempt to exploit, invert, or discredit the generated scenarios. The goal is not only to improve robustness but also to expose hidden assumptions that might otherwise go unchallenged.

Practically, wargames are time-intensive to design, run, and analyze, often requiring significant preparation and coordination. Also, data collection can be uneven, and replicability is a major challenge, especially when games depend heavily on player interaction and improvisation.

9 Conclusion

This paper has explored the intersection of artificial intelligence, scenario planning, and wargaming under the framework of AI-enabled scenario planning (AIESP). In an era of mounting policy complexity, accelerating information flows, and epistemic uncertainty, traditional tools for foresight and decision support have shown their limitations.

The AIESP pipeline represents a significant advancement in the integration of generative AI with epistemic validation mechanisms. Its modular structure, comprising planning, synthesis, generation, and red-team auditing, enables flexibility across use cases. This corruption-aware generative model allows uncertainty to be formally modeled and corrected via Bayesian inference. As more epistemic errors are identified downstream, e.g., through metascience counter-evidence or red-team audits, the model priors can be dynamically updated, improving reliability over time.

The proposed system demonstrates strong potential for scalability. Components such as metascience synthesis and Bayesian correction can be run asynchronously, allowing for distributed processing of high-volume data. The use of counter-evidence retrieval, topic clustering, and rollback-enabled drafting also shows promise for automating large parts of the policy review pipeline without sacrificing quality.

However, this scalability is not without trade-offs. Model complexity introduces opacity, making it difficult for non-technical stakeholders to understand or trust the system's recommendations. Furthermore, computational resource demands, particularly during fine-tuning, inference, and large-scale evidence retrieval, may be prohibitive for smaller organizations or governments with limited infrastructure. Reproducibility also remains a challenge, as variations in prompt construction, LLM behavior, and external data shifts can all affect outputs.

Despite these constraints, the modular design offers a path forward: institutions could adopt only selected modules based on their capacity, or plug in third-party tools (e.g., domain-specific retrieval engines or auditing frameworks). Ultimately, AIESP represents a step toward more reflexive, auditable, and adaptive scenario planning, where AI is not simply a content generator but a participant in structured epistemic dialogue.

Looking forward, the integration of AI scenario generation with wargaming suggests a promising trajectory for reflexive, adaptive policymaking. Rather than providing fixed answers, these tools enable continuous learning: models simulate; agents contest; auditors evaluate; and insights feed back into model correction.

As discussed, challenges remain around model opacity, sponsor bias, anthropomorphizing agents, and reproducibility. Moreover, the dual-use risks of generative AI for deregulation or strategic manipulation must be ethically governed. Nevertheless, the trajectory is clear: the convergence of AI, wargaming, and epistemic validation offers a powerful toolkit for 21st-century policymaking—one that acknowledges complexity, invites contestation, and supports decisions grounded in both strategic imagination and analytical rigor.

A Appendix

B Hypothesis for AI in Wargaming

(1) Representation

Hypothesis 1 (H1: Representational validity)**.** *AI agents trained in high-fidelity wargame representations will show smaller sim-to-real divergence than those trained in abstract environments.*

Hypothesis 2 (M1: Generative scenario fidelity)**.** *LLM-based pipelines fine-tuned on doctrine can produce causally coherent "road-to-war" briefs non-inferior to human versions while cutting preparation time by* ≥ 60

(2) Consequential Decision-Making

Hypothesis 3 (H2: Consequential learning)**.** *AI agents receiving adjudicated, multi-turn payoffs learn more deterrence-stable strategies than agents trained on single-step rewards.*

Hypothesis 4 (M2: Adaptive opposition)**.** *Replacing scripted Red teams with adaptive RL agents increases player-strategy variance by* ≥ 30

(3) Adjudication

Hypothesis 5 (H3: Decision-grounding through adjudication)**.** *Incorporating adjudication feedback loops improves human calibration of risk under uncertainty.*

Hypothesis 6 (M3: Uncertainty-aware adjudication)**.** *Probabilistic neural models outputting full predictive distributions reduce post-game analyst disputes by* ≥ 40

(4) Immersion

Hypothesis 7 (H4: Immersion-induced trust calibration)**.** *Players immersed in high-stakes wargames over-trust AI recommendations when confidence calibration is lowest.*

Hypothesis 8 (M4: Personalised narrative AI)**.** *Dynamic narratives conditioned on psychometrics improve immersion by* ≥ 1 *Likert point.*

(5) Bespoke Design

Hypothesis 9 (H5: Red-team discovery)**.** *Custom-designed games maximising* surprise *uncover rare but critical AI failure modes.*

Hypothesis 10 (M5: Auto-tuned mechanics)**.** *Genetic search over rule-space locates optimal mechanics in* < 500 *simulations with* > 0.8 *correlation to human playtests.*

C Governance Challenges Unique to Generative Policy Support

1. *Synthetic Evidence Floods.* Generative systems can produce unlimited "white-paper" artefacts that *look* authoritative, polluting the information environment that policymakers sample (Seger et al., 2023).
2. *Adversarial Prompt Injection as Lobbying.* Interest groups can optimise prompts that steer LLM policy drafts toward preferred framings, a computational analogue to classic *venue shopping* (Baumgartner et al., 2009).
3. *Model-Drift Feedback Loops.* When generated scenarios influence real-world data collection (e.g., budget priorities), future training data becomes biased, creating an *autophagic* loop (Kasirzadeh, 2023).
4. *Epistemic Responsibility Gap.* Traditional impact assessments assign causal responsibility to modellers; generative pipelines complicate attribution because *no single human* endorses every token (Floridi and Cowls, 2023).

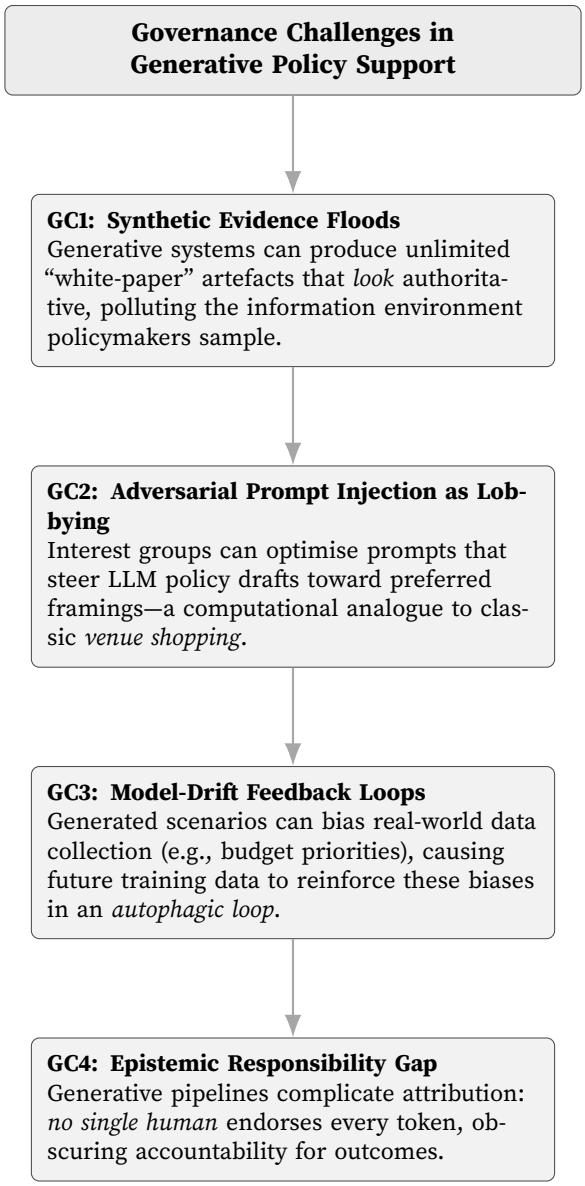


Figure 5 Governance challenges unique to generative policy support.

References

David E Banks. The methodological machinery of wargaming: A path toward discovering wargaming's epistemological foundations. *International Studies Review*, 26(1):viae002, 2024.

Ivanka Barzashka. Wargaming: How to turn a vogue into a science. *Bulletin of the Atomic Scientists*, 2019.

Frank R Baumgartner, Jeffrey M Berry, Marie Hojnacki, et al. Lobbying and policy change: Who wins, who loses, and why. 2009.

Ron Bradfield, George Wright, George Burt, George Cairns, and Kees Van Der Heijden. The origins and evolution of scenario techniques in long range business planning. *Futures*, 37(8):795–812, 2005.

Carl H. Builder. *Toward a Calculus of Scenarios*. RAND Corporation, Santa Monica, CA, 1983.

Thomas Cave and Reuben Binns. Decoding intentions: Artificial intelligence and military decision-making.

Technical report, CSET Georgetown, 2022. URL https://cset.georgetown.edu/publication/decoding-intentions/.

Peter Cornelius, Alexander Van de Putte, and Mattia Romani. Three decades of scenario planning in shell. *California management review*, 48(1):92–109, 2005.

Nicholas Diakopoulos. Algorithmic agenda-setting: How llms embed policy biases. *Digital Policy, Regulation and Governance*, 25(3):215–232, 2023.

Brian D. Earp, Sebastian Porsdam Mann, Mateo Aboy, Edmond Awad, Monika Betzler, Marietjie Botes, Rachel Calcott, Mina Caraccio, Nick Chater, Mark Coeckelbergh, Mihaela Constantinescu, Hossein Dabbagh, Kate Devlin, Xiaojun Ding, Vilius Dranseika, Jim A. C. Everett, Ruiping Fan, Faisal Feroz, Kathryn B. Francis, Cindy Friedman, Orsolya Friedrich, Iason Gabriel, Ivar Hannikainen, Julie Hellmann, Arasj Khodadade Jahrome, Niranjan S. Janardhanan, Paul Jurcys, Andreas Kappes, Maryam Ali Khan, Gordon Kraft-Todd, Maximilian Kroner Dale, Simon M. Laham, Benjamin Lange, Muriel Leuenberger, Jonathan Lewis, Peng Liu, David M. Lyreskog, Matthijs Maas, John McMillan, Emilian Mihailov, Timo Minssen, Joshua Teperowski Monrad, Kathryn Muyskens, Simon Myers, Sven Nyholm, Alexa M. Owen, Anna Puzio, Christopher Register, Madeline G. Reinecke, Adam Safron, Henry Shevlin, Hayate Shimizu, Peter V. Treit, Cristina Voinea, Karen Yan, Anda Zahiu, Renwen Zhang, Hazem Zohny, Walter Sinnott-Armstrong, Ilina Singh, Julian Savulescu, and Margaret S. Clark. Relational norms for human-ai cooperation, 2025. URL https://arxiv.org/abs/2502.12102.

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, and Noam Brown. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. doi: 10.1126/science.ade9097. URL https://www.science.org/doi/abs/10.1126/science.ade9097.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*, 2025.

Luciano Floridi. Ai as agency without intelligence: The case of ai as an epistemic engine. *Minds and Machines*, 32:349–365, 2022.

Luciano Floridi and Josh Cowls. The responsibility gap in ai governance. *Nature Machine Intelligence*, 5: 456–458, 2023.

Jay W Forrester. World dynamics. 1971.

Silvio O Funtowicz and Jerome R Ravetz. Science for the post-normal age. *Futures*, 25(7):739–755, 1993.

Deep Ganguli, Liane Lovitt, Jackson Kernion, et al. Capacity for deception in large language models. *arXiv preprint arXiv:2306.08161*, 2023.

Andrew Gelman, Aki Vehtari, Daniel Simpson, et al. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

Andreas Groll, Thomas Kneib, et al. Variational autoencoders for climate scenario generation. *Environmetrics*, 33(2):e2712, 2022.

Aggie Hirst. Play in(g) international theory. *Review of International Studies*, 45(5):891–914, 2019.

Tom Hope, Jesse Brimhall, Cong Lu, et al. Scim: An open-source metascience engine for real-time evidence surveillance. *Journal of the Association for Information Science and Technology*, 2023.

Michael Horowitz and Gregory Allen. Ai for military decision-making. Technical report, CSET Georgetown, 2021. URL https://cset.georgetown.edu/publication/ai-for-military-decision-making/.

Bin Huang, Peng Zhao, et al. Diffusion models for urban scenario planning. *Computers, Environment and Urban Systems*, 98:101912, 2023.

Marcus Hughes, Yoni Halpern, et al. Narrative elaboration with large language models for policy exploration. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 412–423, 2023.

Atoosa Kasirzadeh. Autophagy of the epistemic commons: Generative ai and the risks of knowledge collapse. *AI & Society*, 2023.

Hannah Kirk, Yohan Jun, Luca Volpin, Maarten Sap, et al. Bias out-of-the-box: How rlhf can amplify demographic stereotypes. *arXiv preprint arXiv:2305.13856*, 2023.

Robert J Lempert, Steven W Popper, and Steven C Bankes. Shaping the next one hundred years: New methods for quantitative, long-term policy analysis. 2003.

Erik Lin-Greenberg. Wargame of drones: Remotely piloted aircraft and crisis escalation. *Journal of Conflict Resolution*, 66(10):1737–1765, 2022.

Erik Lin-Greenberg, Reid BC Pauly, and Jacquelyn G Schneider. Wargaming for international relations research. *European Journal of International Relations*, 28(1):83–109, 2022.

Tshilidzi Marwala. *Rational Machines and Artificial Intelligence*. Elsevier, Netherlands, January 2021. doi: 10.1016/B978-0-12-820676-8.09990-7. Publisher Copyright: © 2021 Elsevier Inc. All rights reserved.

Tshilidzi Marwala and Evan Hurwitz. Artificial intelligence and economic theories, 2017. URL https://arxiv.org/abs/1703.06597.

Roger C Mason. Wargaming: its history and future. *The International Journal of Intelligence, Security, and Public Affairs*, 20(2):77–101, 2018.

Dennis Meadows and Jorgan Randers. *The limits to growth: the 30-year update*. Routledge, 2012.

Joshua M Nicholson et al. Metascience as a service: How large-scale analytics can detect emerging research trends. *PLOS Biology*, 20(10):e3001812, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Cai, et al. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

Kenneth Payne. *I, Warbot: The Dawn of Artificially Intelligence Conflict*. Hurst & Company, July 2021. ISBN 1787384624.

Peter P Perla. *The Art of Wargaming*. Naval Institute Press, 1990.

George P Richardson. System dynamics: Simulation for policy analysis from a feedback perspective. In *Qualitative simulation modeling and analysis*, pages 144–169. Springer, 1991.

Robert C Rubel. The epistemology of war gaming. *Naval War College Review*, 59(2):108–128, 2006.

Paul J H Schoemaker. Scenario planning: a tool for strategic thinking. *Sloan Management Review*, 36(2): 25–40, 1995.

Todd Sechser and Savannah Goldman. Ai and nuclear stability. *Journal of Strategic Studies*, 44(7):937–966, 2021.

Elizabeth Seger, Shahar Avin, et al. Hallucinating white papers: Synthetic evidence and the governance risks of generative ai. *arXiv preprint arXiv:2306.05866*, 2023.

Ying Shao, Yang Liu, Zhe Zhang, et al. Gamellm: A multi-agent wargaming framework with large language models. *arXiv preprint arXiv:2307.11346*, 2023.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value profiles for encoding human variation, 2025. URL https://arxiv.org/abs/2503.15484.

Christopher Summerfield, Lisa Argyle, Michiel Bakker, Teddy Collins, Esin Durmus, Tyna Eloundou, Iason Gabriel, Deep Ganguli, Kobi Hackenburg, Gillian Hadfield, Luke Hewitt, Saffron Huang, Helene Landemore, Nahema Marchal, Aviv Ovadya, Ariel Procaccia, Mathias Risse, Bruce Schneier, Elizabeth Seger, Divya Siddarth, Henrik Skaug Sætra, MH Tessler, and Matthew Botvinick. How will advanced ai systems impact democracy?, 2024. URL https://arxiv.org/abs/2409.06729.

Graham M Turner. A comparison of the limits to growth with 30 years of reality. *Global environmental change*, 18(3):397–411, 2008.