Check for updates

AI & Innovation

WILEY | IDEAS

PERSPECTIVE ARTICLE OPEN ACCESS

Playing in the Dark: Invisible Chess as a Laboratory for Strategic AI

Paolo Ciancarini

Università di Bologna, Bologna, Italy

Correspondence: Paolo Ciancarini (paolo.ciancarini@unibo.it)

Received: 14 March 2026 | **Revised:** 6 July 2026 | **Accepted:** 9 July 2026

ABSTRACT

Artificial intelligence research has historically relied on games such as chess as benchmarks for progress in strategic reasoning. While chess and Go domains have produced remarkable advances—from Deep Blue to AlphaZero—they share a fundamental property that limits their relevance for many real-world applications: perfect information. In contrast, most strategic environments in domains such as international relations are characterized by incomplete, uncertain, and asymmetric information. This perspective article makes three contributions. First, it proposes a structured taxonomy of imperfect-information chess variants—including Kriegspiel and related "invisible board" environments—organized by the type and degree of information asymmetry they impose. Second, it presents an empirically grounded case study of Darkboard, the strongest computer Kriegspiel player to date, whose decade-long record on the Internet Chess Club and three Computer Olympiad gold medals provides concrete evidence of both the power and the structural limits of AI under adversarial uncertainty. Third, it derives governance implications for the design and evaluation of AI systems operating in adversarial real-world environments, where an opponent can model and exploit the inference procedure of an AI agent. We examine how the digital transformation of chess has shaped the development of AI methods and highlight the epistemic limitations of perfect-information benchmarks. We discuss how games with imperfect information introduce challenges central to modern AI systems, including belief-state reasoning, partial observability, and adversarial uncertainty. Finally, we outline implications for the future of AI innovation, suggesting that imperfect-information strategic games may serve as valuable interdisciplinary laboratories for exploring the design, governance, and evaluation of intelligent systems operating under uncertainty.

1 | Introduction

Strategic decision-making has always been shaped by the tools available to decision-makers, politicians, and analysts. The digital transformation of conflicts, accelerated by advances in artificial intelligence (AI), autonomous systems, and networked sensing, is transforming the tempo, scale, and epistemic conditions of international relations and conflicts at a pace that outstrips established doctrine [1]. Autonomous weapons, AI-curated intelligence feeds, cyber operations, and generative disinformation campaigns do not merely add new instruments to the strategic toolkit—they alter the fundamental information environment in which decisions are made, compressing timelines, multiplying actors, and deepening the adversarial uncertainty rather than dispersing it [2].

Consider a concrete case. A cyber defense system must decide, in near-real time, whether an anomalous network event represents a genuine intrusion, a probe, or a deliberate false flag designed to trigger a defensive response that itself reveals some protected infrastructure. The system observes only fragments of the adversary's behavior; the adversary knows this and acts accordingly. No perfect-information benchmark—no matter how combinatorially complex—can simulate, train or evaluate an AI system for this epistemic situation. This paper argues that

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *AI & Innovation* published by John Wiley & Sons Australia, Ltd on behalf of Institute for Digital Economy and Artificial Systems (IDEAS).

strategy games based on invisible moves and negative information, like imperfect-information chess variants, and Kriegspiel in particular, provide precisely the structured laboratory that is missing.

A recurring trope in discussions of AI-augmented strategy is the game of chess. The milestones are well known: Garry Kasparov's defeat by IBM Deep Blue in 1997 [3], the rise of open-source engines such as Stockfish [4], and the self-taught mastery demonstrated by AlphaZero in 2017 [5] are routinely invoked as proof that AI has conquered strategic reasoning. The implication, sometimes stated explicitly, is that the same algorithmic breakthroughs will soon dominate the strategic domains of diplomacy, defense, and international security. Yet chess is a game of perfect information: both players see the entire board and the opponent's moves at all times. Every piece is known; every move is seen. Only the optimal response is not. International relations and strategic decisions in the real-world share almost none of these properties. Positions are hidden, intelligence is partial and often deliberately falsified, and the "rules" can change mid-game. Analogies drawn from classical chess therefore risk importing a fundamentally misleading epistemology into the analysis of AI-augmented conflict [1].

This paper proposes a corrective. We argue that invisible chess—a class of game variants of the Kriegspiel wargame invented in Prussia in the early nineteenth century [6] and formalized as a two-player partial-information game [7]—provides a substantially more accurate analog of digitally transformed strategic environments than its perfect-information counterpart. In the variant named Kriegspiel chess, each player sees only their own pieces; the position and moves of the opponent are hidden and must be inferred through attempted moves, probe captures, and terse responses from an umpire. The resulting epistemic situation—persistent uncertainty, active deception, belief-state management, and decisions under irremediable incompleteness of information—mirrors the structural properties of modern conflict with a fidelity that standard chess simply cannot offer [8]. We use invisible chess as the umbrella term for this class of variants; the terms imperfect-information chess and partial-information chess are used in the literature interchangeably, but we prefer imperfect information as it aligns with standard game-theoretic usage [7].

Our guiding question is: Can invisible chess become a superior benchmark for AI under strategic uncertainty?

Our argument proceeds in four steps. First, we shortly recall the digital transformation of chess, from early computer programs through the engine revolution to neural network-based self-play, highlighting what this trajectory reveals about AI capabilities and their limits when the information environment is idealized. Second, we introduce Kriegspiel as a domain of structured strategic uncertainty, reviewing the combinatorial challenges that distinguish it from standard chess and make it resistant to the algorithmic techniques that proved decisive in the perfect-information case. Third, we draw on the development and competitive record of Darkboard—one of the strongest invisible chess-playing programs produced to date, developed over several years at the University of Bologna—to ground our theoretical claims in concrete empirical evidence about where AI succeeds and fails under uncertainty [8]. Fourth, we transpose these findings onto the landscape of international relations and AI-augmented conflicts, showing that the mechanisms that confound invisible chess-playing engines—hidden state, adversarial deception, and irreducible epistemic uncertainty—are not obstacles that digital transformation removes but dynamics that it amplifies.

The analysis carries three implications for the governance of AI-augmented strategy. The first concerns the centaur model of decision-making: Just as human–AI teams have proven more robust than either partner alone in Kriegspiel and related imperfect-information games [9], effective command in uncertain digital environments will require sustained human judgment at precisely the points where AI confidence is lowest. The second concerns digital sovereignty: Nations whose strategic AI systems are trained primarily on perfect-information benchmarks may be investing in brittle capabilities that fail under adversarial uncertainty, with important implications for the current wave of national AI strategies [10]. The third concerns the governance of autonomous weapons: The limits we identify in invisible chess AI are structural, not merely technical, and argue against delegating irreversible lethal decisions to systems that cannot resolve fundamental uncertainty about hidden state. We use autonomous weapons throughout; the terms lethal autonomous weapons systems (LAWS) and autonomous lethal engagement appear in the literature but are not used here except where directly cited.

1.1 | A Note on Contribution Type

This is primarily an integrative perspective article: It does not report new experimental results or introduce a new algorithm, but synthesizes findings from game AI, security studies, and AI governance through a novel conceptual lens—the choice of benchmarking strategic games under imperfect information. The taxonomy and case study serve as empirical anchors; the governance argument is the intended secondary payoff.

The remainder of the paper is organized as follows. Section 2 reviews the digital transformation of chess and its limitations as a strategic analogy. Section 3 introduces Kriegspiel (chess under uncertainty) and its computational complexity. Section 4 presents Darkboard as a case study in AI under imperfect information. Section 5 develops the parallels with AI-augmented warfare and international relations. Section 6 draws out implications for AI innovation, digital sovereignty, and autonomous systems governance. Section 7 presents the conclusion.

2 | The Digital Transformation of Chess

Chess has served as a laboratory for artificial intelligence research since the field's inception. Shannon's foundational paper on computer chess [11] inaugurated the study of the game of chess as a canonical testbed for machine reasoning, and the subsequent 7 decades of progress—from early alpha-beta pruning through dedicated hardware to neural self-play—chart the arc of AI development as faithfully as any benchmark [3]. Understanding what this trajectory reveals, and what it conceals, is essential before drawing any analogy between chess mastery and strategic competence in the real world.

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.1 | The Eras of Computer Chess

The history of computer chess can be divided into three overlapping eras, each characterized by a different dominant paradigm.

2.1.1 | Era 1: Handcrafted Search (1950–1997)

The first era was defined by minimax search with alpha-beta pruning, combined with hand-tuned evaluation functions encoding grandmaster knowledge [3]. Programs such as Chess 4.5, Cray Blitz, Deep Thought, and ultimately Deep Blue occupied this paradigm. Deep Blue's victory over Garry Kasparov in the 1997 match marked the symbolic endpoint of the era: A purpose-built machine, capable of evaluating up to 200 million positions per second, had surpassed the world's strongest human player [3]. The result was widely reported as a watershed in AI, yet the system was brittle by modern standards—dependent on an enormous database of grandmaster games, manually tuned evaluation weights, and specialized hardware that could not be repurposed for any other task.

2.1.2 | Era 2: Open-Source Engine Dominance (1997–2017)

The second era saw the rise of open-source engines—most notably Stockfish—running on commodity hardware [4]. Improvements in search efficiency, deeper endgame tablebases, and increasingly refined evaluation functions have pushed engine strength well beyond that of any human. By the early 2010s, even smartphone applications were routinely defeating grandmasters. Crucially, this era also produced the centaur or advanced chess concept, pioneered by Kasparov himself: human–computer teams competing against other human–computer teams, demonstrating that the combination of human intuition and machine calculation could outperform either partner in isolation [9]. This finding—that complementarity rather than replacement may be the optimal human–AI configuration—will recur as a central theme in our analysis of strategic environments.

2.1.3 | Era 3: Neural Self-Play (2017–Present)

The third era was inaugurated by AlphaZero, DeepMind's general reinforcement learning system, which learned chess entirely through self-play—without access to human games—and within 24 hours surpassed Stockfish by a substantial margin [5]. AlphaZero replaced handcrafted evaluation with a deep neural network trained end-to-end and exhaustive search with policy-guided Monte Carlo Tree Search (MCTS). Its successor, Stockfish 16 with NNUE (Efficiently Updatable Neural Network) evaluation, represents the current state of the art: A hybrid of classical search and neural evaluation that achieves an estimated Elo rating above 3500, more than 600 Elo points above the strongest human player. For all practical purposes, perfect-information chess is a solved problem at the level of play relevant to human competition.

2.2 | What the Chess AI Trajectory Demonstrates

The chess AI story yields several genuine insights that transfer to broader discussions of AI and strategy. First, it demonstrates that search and evaluation at superhuman scale are achievable: Given a well-defined state space and a reliable evaluation signal, deep learning and tree search can jointly master extraordinarily complex combinatorial problems. Second, it validates the value of self-play as a training paradigm: AlphaZero required no labeled human data, suggesting that AI can develop strategic competence in domains where expert knowledge is scarce or unreliable [5]. Third, the centaur evidence suggests that human–AI complementarity is robust even when AI is nominally superior, a finding with direct implications for command and decision-support system design [9].

2.3 | The Fundamental Limitation: Perfect Information

Despite these contributions, chess is a game of perfect information [7]: Both players observe the complete board at all times, so uncertainty is purely computational (evaluating an observable tree) rather than epistemic (ignorance of the current state). Real strategic environments—military operations, cyber campaigns, geopolitical crises—violate this property fundamentally: positions are hidden, intelligence is falsified, and the rules themselves can change. An AI optimized for perfect-information play also inherits the closed-world assumption (fixed rules, known capabilities, and unambiguous objectives), a brittleness well documented in adversarial machine learning [12] and increasingly recognized in autonomous systems research.

2.4 | The Need for an Imperfect-Information Analog

These limitations do not invalidate chess as a partial analogy for strategy—its combinatorial depth, long-horizon planning requirements, and adversarial structure remain instructive. However, they motivate the search for a richer analog that preserves these virtues while incorporating genuine epistemic uncertainty. We argue in the following section that Kriegspiel—chess played with hidden opponent pieces—provides precisely this bridge, and that the history of computer Kriegspiel illuminates the structural challenges that AI systems will face in any domain where the state of the world cannot be fully observed—a challenge we examine in the next section through the lens of Kriegspiel and its modern descendants.

3 | Chess Under Uncertainty: A Taxonomy

This section introduces three classes of imperfect-information chess variants—referee-mediated (Kriegspiel), passive-visibility (Dark Chess and Fog of War), and budget-allocated sensing (RBC)—compares their information structures and AI challenge profiles and closes with a unified comparative framework (Table 1).

The limitations identified in Section 2 motivate a family of chess variants in which information about the game state is withheld from one or both players. Each player must reason over a belief state—a probability distribution across board positions consistent with their observations—rather than a single known position. The variants differ in the mechanism by which imperfect information is revealed, and this difference determines the AI

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 | Invisible chess variants.

Variant	Mechanism	Visibility	Feedback	AI challenge	Strategic analog
Kriegspiel	Referee-mediated	Own pieces only	Umpire: Illegal moves, captures, check direction	Belief-state tracking; probe sequencing; deception	Fog of war; HUMINT; limited ISR
Dark chess	Passive geometric	Immediate capture zone only	None	Near-total uncertainty; risk-sensitive planning; probabilistic inference	Denied/degraded environments; electronic warfare; sensor-spoofed environment
FoW chess	Passive geometric	Own threat radius	None (passive only)	Mobility–visibility trade-off; proactive reconnaissance; superhuman AI via general search [13]	Sensor-based ISR; drone coverage; cyber situational awareness
RBC	Budget-allocated	Explicit 3 × 3 sensing window	Truthful within sensed region; silent illegal moves	Sensing budget allocation; neural information-set weighting [14]; DSMCP [15]	ISR collection planning; finite surveillance asset allocation

challenges each imposes. We distinguish three classes: (i) referee-mediated feedback (Kriegspiel), (ii) passive geometric visibility (Dark Chess and Fog of War), and (iii) budget-allocated active sensing (RBC).

3.1 | Kriegspiel: Referee-Mediated Information

Kriegspiel was invented by Georg von Reisswitz Jr. in Prussia in 1824 as a military training tool, subsequently adopted by Prussian—and later many other—general staff as a wargaming educational instrument [6]. Its chess variant was practiced by members of the RAND corporation [16] of the caliber of John Nash, Lloyd Shapley, and John D. Williams [17]. Each player sees only their own pieces; the opponent's pieces and moves are hidden. A neutral umpire adjudicates attempted moves: An illegal move prompts "*Illegal move*", forcing a retry and revealing information through the pattern of failures. Captures are announced ("Capture on e4") without identifying the capturing piece. Check is announced by direction ("Check on the file" and "Check on the diagonal") but not by piece.

This creates a partially observable two-player zero-sum game with imperfect information and perfect recall. The belief state can be astronomically large. Sakuta and Iida introduced the metaposition to manage this complexity: a compact encoding of all opponent configurations consistent with observations so far, updated incrementally as umpire responses arrive [18]. Standard alpha-beta pruning is inapplicable—the agent must reason over a distribution of trees, each corresponding to a different possible true board state.

The key AI capabilities required are qualitatively different from those that conquered standard chess: (i) *belief-state tracking*—maintaining a probabilistic model of hidden state under ambiguous observations; (ii) *robust action selection*—choosing moves that perform well across the belief distribution rather than optimally in a single state; (iii) *information gathering*—probing to reduce uncertainty at the cost of tempo; and (iv) *deception*—misleading the opponent about one's own position. These capacities map directly onto the epistemic challenges of AI-augmented strategic environments, as we develop in Section 5.

3.2 | Passive Visibility: Dark Chess and Fog of War

A second class dispenses with the referee entirely. Visibility is determined geometrically by piece placement: The agent passively observes whatever falls within its visibility radius, with no external feedback and no cost attached to observation. The two main variants differ in how wide that radius is.

3.2.1 | *Dark Chess*

In Dark Chess (Dunkelschach) [19], each player sees only the squares occupied by their own pieces and the squares those pieces could immediately capture—a strictly local radius with no projection into deeper squares. There is no umpire and no illegal-move feedback. The result is the most information-constrained of the variants considered here: A player has no knowledge of the opponent's formation beyond the immediate contact zone. Dark Chess should not be confused with Chinese Dark Chess (Banqi), which is played on a 4 × 8 grid with Xiangqi pieces, introduces a stochastic element through random face-down piece revelation, and constitutes a distinct game-theoretic problem.

3.2.2 | *Fog of War Chess*

Fog of War (FoW) chess widens the visibility radius: Each player sees all squares that their own pieces could legally move to, illuminated by the threat radius of one's own forces [13]. Everything outside that radius is hidden. Unlike Kriegspiel, there is no umpire and no penalty for attempting a move—the only way to observe a region is to move a piece into a position that threatens it.

This mechanic introduces a fundamental trade-off between mobility (moving pieces to strong attacking or defensive squares) and visibility (positioning pieces to maximize board coverage). In Kriegspiel, information is revealed reactively through failed move attempts; in FoW, it is controlled

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

proactively through piece placement. The trade-off directly mirrors the intelligence-maneuver tension in military operations: Committing forces to an objective reduces their availability for reconnaissance, whereas dispersing them for surveillance reduces striking power. The recent achievement of superhuman performance in FoW by Obscuro [13] demonstrates that principled general search can overcome even this demanding information structure.

3.3 | Reconnaissance Blind Chess: Sensing on Budget

The third class makes information acquisition an explicit first-class decision. In Reconnaissance Blind Chess (RBC), adopted as a competitive AI benchmark through a series of NeurIPS challenge competitions [20], a player may sense any 3×3 region of the board on each turn, receiving a complete and truthful report of the pieces in that window; outside it, the board is fully hidden. The player then makes a standard chess move, handled silently if illegal—the piece remains in place with no umpire feedback.

The central strategic challenge is the allocation of the sensing budget: which region to observe each turn, given that unobserved regions remain hidden and the opponent may have moved since the last observation. This maps directly onto the intelligence, surveillance, and reconnaissance (ISR) collection problem in defense operations, where a commander with finite surveillance assets must decide which sectors to cover on each collection cycle, accepting persistent uncertainty elsewhere.

Two state-of-the-art AI approaches have addressed this challenge. Bertram and others [14] developed a neural information-set weighting approach: rather than treating all configurations consistent with observations as equally likely, their system learns a posterior distribution over opponent positions and concentrates sensing on regions where uncertainty is strategically consequential, demonstrating meaningful improvements over uniform-belief baselines. Clark [15] proposed deep synoptic Monte Carlo planning (DSMCP), the system underlying the winning Penumbra agent in the 2020 RBC Leaderboard Challenge. DSMCP represents the belief state as a particle filter over sampled opponent configurations, encodes each particle via a permutation-invariant synopsis, and allocates the sensing budget through Upper Confidence bounds applied to Trees (UCT) bandit planning that maximizes expected information gain. Compared with Darkboard's constraint-based meta-position (Section 4), the particle-filter approach scales more gracefully in high-entropy middlegames, though both architectures remain vulnerable to adversaries who deliberately engineer high-entropy belief states.

3.4 | Comparative Framework

Table 1 summarizes the information structure, AI challenge profile, and strategic analog of the four variants, organized by information acquisition mechanism.

The three mechanisms constitute a progression in the agency the player has over information acquisition. In Kriegspiel, information arrives as an uninvited by-product of action—the umpire speaks only when spoken to, and only in response to moves that were attempted anyway. In Dark Chess and FoW, information is a passive consequence of position—the agent controls what it sees only indirectly, through where it chooses to stand. In RBC, information is a deliberate purchase—the agent explicitly decides what to observe on every turn, making sensing allocation a primary strategic variable on par with move selection.

This progression also maps onto increasing operational relevance for AI-augmented defense systems: Kriegspiel mirrors human intelligence (HUMINT)-dominated environments where information is acquired reactively; FoW and Dark Chess mirror sensor-based ISR where coverage is determined by platform deployment; and RBC mirrors modern collection management where finite ISR assets are actively tasked against prioritized targets. Together, the four variants provide a richer and more complete benchmark family for AI innovation under uncertainty than any single variant alone. In the next section, we examine Darkboard, a strong Kriegspiel engine, as a concrete case study in what is achievable—and what remains out of reach—when AI is confronted with this class of problems.

4 | A Case Study in AI Under Uncertainty

This section grounds the theoretical arguments of Section 3 in a concrete system: it describes Darkboard's technical design (Sections 4–4.3), presents its empirical record, and identifies the specific failure modes that support the governance argument of Section 6.

The theoretical arguments developed in the preceding sections—that imperfect-information chess variants impose qualitatively different computational and epistemic demands on AI systems, and that these demands mirror the structural properties of AI-augmented strategic environments—acquire their fullest force when grounded in a concrete system. Darkboard, a Kriegspiel-playing program developed over several years at the University of Bologna, provides precisely this grounding. In documenting what Darkboard achieves, where it fails, and why, we obtain a set of empirically derived propositions about the limits of AI under uncertainty that transfer directly to the warfare and international relations context developed in Section 5.

4.1 | Origins and Motivation

The Darkboard project originated from a recognition that Kriegspiel occupies a unique position in the landscape of AI-tractable games [21]. Unlike poker or bridge—the other canonical imperfect-information game benchmarks—Kriegspiel retains the full combinatorial depth of chess while imposing genuine epistemic uncertainty about the current board state. This combination makes it simultaneously harder than poker (the underlying game tree is vastly larger) and more strategically rich than simpler hidden-information games. From an AI research perspective, it constitutes an ideal stress test: Any technique that works well in Kriegspiel must solve both the combinatorial planning problem and the belief-state management problem simultaneously, without reducing one to the other.

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Early computer Kriegspiel programs—including entries to the Computer Olympiad in the early 2000s—relied on simple heuristics: treat the most likely opponent position as if it were certain, apply standard chess evaluation, and hope that the resulting play was robust to the true hidden state [22]. This approach, sometimes called most-likely-state (MLS) play, produced programs that performed adequately in simple positions but collapsed in tactically sharp situations where the hidden opponent configuration was decisive. The Darkboard project set out to replace MLS heuristics with principled belief-state reasoning, accepting the computational costs this entailed in exchange for more robust and theoretically grounded decision-making [8, 21].

4.2 | Key Technical Contributions

4.2.1 | The Metaposition Representation

The central technical innovation of Darkboard is the metaposition: a compact data structure that encodes the set of all board positions consistent with the agent's observation history—its own piece locations, the sequence of umpire responses received, and the logical constraints that follow from them [21]. Rather than maintaining an explicit list of possible board states (which grows exponentially with game length), a metaposition represents the consistent set implicitly through a set of constraints on the squares that opponent pieces can and cannot occupy. Updates are performed incrementally: each new umpire response ("Illegal move", "Capture on e4", "Check on the diagonal") is translated into a set of constraint updates that prune inconsistent configurations from the metaposition. This representation makes belief-state tracking tractable in the opening and middlegame, where the constraint network is still manageable, though it remains exponentially hard in the worst case [8].

4.2.2 | Search Under Uncertainty

Standard alpha-beta minimax search is inapplicable in Kriegspiel because there is no single game tree to search: the correct tree depends on the true board state, which is unknown. Darkboard addresses this by adapting a maximax search over sampled positions drawn from the current metaposition. Candidate moves are evaluated by sampling a set of consistent opponent configurations, running a shallow alpha-beta search in each, and aggregating the results. Moves that perform well across the distribution of sampled states are preferred over moves that are optimal for the single most likely state but catastrophic for others. This approach instantiates a form of robust decision-making under uncertainty: optimizing for average-case performance across the belief state rather than best-case performance in a single scenario—a principle with direct applications in military planning under incomplete intelligence [2].

4.2.3 | Probe Sequencing and the Reconnaissance Trade-Off

A distinctive feature of Kriegspiel strategy is the availability of pawn capture attempts as reconnaissance probes: a player may attempt a diagonal pawn capture to discover whether an opponent piece occupies the target square, at the cost of a move tempo if the attempt fails. Darkboard incorporates an explicit model of probe value: the expected reduction in belief-state entropy achieved by a probe attempt, weighed against the positional cost of the tempo [8]. When the metaposition is highly uncertain and a probe can substantially reduce it, Darkboard invests in reconnaissance; when the positional situation is urgent, it exploits its current knowledge. This information-gathering versus exploitation trade-off is one of the most direct structural parallels between Kriegspiel AI and intelligence-driven military operations, where committing forces to a reconnaissance mission diverts them from the main effort.

4.2.4 | Endgame Handling

Some Kriegspiel endgames (e.g., King-Queen vs. King (KQ vs. K), King-Rook versus King (KR vs. K), King-Bishop-Bishop versus King (KBB vs. K), King-Bishop-Knight versus King (KBN vs. K)) were solved to completion using retrograde analysis over the full belief-state space [23]. The corresponding algorithms were included in Darkboard. These results demonstrate that brute-force belief-state search is tractable in constrained endgame settings, where the number of consistent configurations is small enough to enumerate explicitly. AI can achieve full mastery of an imperfect-information domain when the belief state is sufficiently constrained.

4.3 | Performance and Empirical Results

We tested the Darkboard program against human opponents using the Internet Chess Club (ICC). This server offered a Kriegspiel game under the name of "Wild 16″, where the player community comprised several hundred humans and a handful of programs. The principal empirical record derives from ICC where Darkboard accumulated about 10,000 games across two versions (Darkboard 1.0: $N = 2,442$; Darkboard 2.0: $N = 7,121$) against human and computer opponents over nearly five years [24]. Moreover, Darkboard competed in the International Computer Games Association (ICGA) Computer Olympiad Kriegspiel tournament across three editions (2006, 2009, 2024), winning all of them.

4.3.1 | Performance Statistics

Table 2 summarizes Darkboard's overall ICC record. With an average score of 0.5 across approximately 10,000 games and a corrected average Elo of ∼1,700, Darkboard played at an above-average human level; notably, about 40% of its games were played against the top-20 ICC players, against whom its average score dropped sharply to 0.26 [24]—direct empirical evidence for the structural human advantage discussed in Section 4.4.

Against computer opponents, tested primarily at the ICGA Computer Olympiad, Darkboard's advantage was sharpest when belief-state management was decisive: MLS-based opponents were consistently vulnerable to configurations in which the most probable state was a deliberate decoy. In high-entropy middlegames, outcomes shifted closer to chance, consistent with the fundamental intractability of the general case [8].

4.3.2 | Two Illustrative Adversarial Cases

Two recurring patterns illustrate how skilled human players exploit Darkboard's inference procedure.

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 | Darkboard performance summary.

Metric	Darkboard
Games played	~10,000
Unique opponents	~600 (~17 games each)
Avg. opponent Elo (corrected)	~1700
Avg. score	0.5
% Games versus higher ELo	~57%
Games versus top-20 players	~4000 (40%)
Average score versus top-20 players (points/game)	0.26

Case 1: False-structure injection. Expert opponents systematically constructed pawn formations that, under Darkboard's constraint-propagation update rules, assigned high probability to a mirror-image of the true configuration. A tactical combination on the under-weighted side then produced decisive material gain—a direct analog of feeding false order-of-battle data to an adversary's intelligence fusion system.

Case 2: Probe-sequence baiting. Darkboard invests in reconnaissance when the expected entropy reduction exceeds a positional cost threshold [8]. Expert opponents created positions where initiating a probe was locally attractive but strategically counterproductive: the tempo cost exposed the king-side structure, which was then exploited. This mirrors the ISR allocation problem in which an adversary models the collection-management algorithm to trigger a predictable reallocation.

Both cases share a structural feature: the adversary's advantage derived not from superior chess strength but from a model of Darkboard's inference procedure—a finding with direct implications for AI decision-support design in adversarial environments, discussed in Section 5 [8].

4.4 | Some AI Limits in Uncertain Domains

4.4.1 | *Belief-State Explosion as a Structural Obstacle*

The most important lesson from Darkboard is that the difficulty of imperfect-information play is not merely quantitative—a larger search space requiring more computation—but qualitative: It reflects a structural property of the problem that no increase in computational power straightforwardly resolves.

Remark (Structural intractability of belief-state reasoning). Exact Bayesian belief update over $|S|$ world states requires $O(|S|^2)$ operations per step [25]; in Kriegspiel specifically, the belief state has been observed to exceed 10^{10} physical configurations in practice [26]. Finite-horizon POMDP planning is PSPACE-complete [27, 28]; in the two-agent adversarial setting, optimal planning is NEXP-complete [29]; with unbounded horizons the problem is undecidable in general [30]. These results confirm that the obstacles Darkboard faces are structural, not merely engineering limitations of the current implementation.

As the game progresses and the umpire's responses accumulate, the metaposition can either contract (as constraints rule out configurations) or expand (as new pieces of possible hidden state are introduced by pawn promotions, undiscovered piece movements, or ambiguous captures). In the worst case, the belief state grows faster than it can be pruned, and Darkboard's sampling approximations become increasingly unreliable. This is not an engineering failure; it is a consequence of the combinatorial structure of the problem, analogous to the curse of dimensionality in multi-agent partially observable systems [29].

4.4.2 | *The Limits of Self-Play in Imperfect-Information Domains*

AlphaZero's self-play paradigm succeeded in perfect-information chess because both agents observe the complete game state, enabling unambiguous credit assignment during training: a move that led to a loss was demonstrably suboptimal given the known board [5]. In Kriegspiel, self-play faces a fundamental difficulty: the agent that lost may have played optimally given its observations, and the loss may have been due to an irreducibly unfavorable hidden configuration rather than a strategic error. This complicates credit assignment and slows policy improvement. Recent experiments with MCTS-based approaches to Kriegspiel confirm that neural self-play methods achieve significantly lower performance relative to their perfect-information counterparts, and that the gap widens as the game length increases and the belief state becomes more complex [8]. Closing this gap would require fundamentally new training paradigms—perhaps combining belief-state search with opponent modeling and explicit reasoning about information asymmetry—that remain open research problems.

4.4.3 | *The Irreducible Human Edge*

The most analytically significant finding from Darkboard's game record is the persistence of a human advantage in positions characterized by adversarially maintained uncertainty: situations where a skilled opponent deliberately constructs and sustains a high-entropy belief state, exploiting Darkboard's known tendencies and the statistical regularities of its sampling procedure. Some human players use strategies of "feeding the algorithm false structure"—creating patterns of umpire responses that trigger systematic biases in metaposition updates, leading the program to assign high probability to false configurations and make correspondingly poor decisions [24]. This form of adversarial deception against an AI system is not a theoretical curiosity; it is a direct analog of the information operations, deception campaigns, and sensor spoofing that characterize modern digital conflict [2]. The implication is sobering: the more transparent an AI system's decision procedure, the more exploitable it is by an adversary who can model that procedure and design observations to manipulate it.

4.4.4 | *Prospects for Modern Techniques*

It is natural to ask whether techniques developed since Darkboard's primary development period—MuZero-style model-based reinforcement learning [5], NNUE neural evaluation adapted for imperfect information, or large language model-based intuition for opponent modeling—could substantially close the remaining gap. Our assessment is cautiously negative for the core challenge. These techniques improve efficiency of search and

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

quality of evaluation in tractable positions, but they do not address the structural intractability of belief-state explosion in the general case, nor do they provide a principled solution to adversarial deception against the AI's own inference procedure. They may produce stronger Kriegspiel programs, but they are unlikely to achieve the same degree of dominance that AlphaZero achieved in standard chess—and the reasons why are precisely the reasons that AI systems in real strategic environments should not be expected to achieve reliable autonomous decision-making under adversarially maintained uncertainty.

4.5 | Summary

The Darkboard case study establishes three propositions that support the paper's policy argument. First, AI can achieve strong and in some cases optimal performance in imperfect-information domains when the belief state is locally tractable—constrained by accumulated observations to a manageable set of consistent states. Second, AI performance degrades systematically and, in the worst case, catastrophically, when belief-state complexity is globally and adversarially maintained—when the opponent actively works to keep the agent uncertain. Third, skilled human players retain a meaningful advantage precisely in this second regime, through forms of deception and opponent modeling that exploit the statistical regularities of AI inference procedures. These three propositions carry direct implications for the design and governance of AI-augmented strategic systems, which we develop in Section 5.

5 | Uncertainty in Strategic Environments

This section translates the Darkboard propositions into three real-world domains—digital conflict and the OODA cycle, cyber operations, and AI in international relations—showing in each case how adversarial epistemic uncertainty produces the same structural failure modes identified in the game laboratory.

The propositions established by our case study describe a structural feature of intelligent decision-making under imperfect information that recurs across domains where adversaries actively shape each other's epistemic state.

5.1 | Digital Transformation Amplifies Uncertainty

A common assumption is that digital transformation reduces uncertainty through more sensors and faster data fusion [2]. While digital systems increase the volume of available data, they simultaneously expand the adversarial surface across which an opponent can introduce false or misleading information: some examples are GPS spoofing, radar deception, deepfake video, and adversarial perturbations of AI classification systems are forms of digitally enabled manipulation with no meaningful pre-digital analog [1]. The net effect is not a reduction in uncertainty but a transformation of its character: adversarial uncertainty replaces simple absence of information. This is precisely the condition identified as the hardest—not the absence of observations, but the presence of observations designed to mislead. An AI system that cycles through its sense-decide-act loop at machine speed on the basis of adversarially manipulated inputs will reach confidently wrong decisions faster than a slower, more skeptical human analyst [12].

5.2 | The Decision Cycle and Belief-State Management

John Boyd's Observe–Orient–Decide–Act (OODA) framework characterizes decision tempo in competitive environments [31]. Proponents of AI-augmented decision-making argue that AI accelerates the Observe and Orient phases by processing information faster than human analysts. The Darkboard evidence qualifies this argument sharply: acceleration is only beneficial when the observations feeding the Orient phase are reliable. When an adversary is actively curating those observations—as skilled Kriegspiel players did against Darkboard—faster processing of corrupted inputs produces faster convergence on false beliefs. The Orient phase requires not merely fast belief updating but meta-cognitive awareness of the reliability of the update process itself: a capacity that current AI architectures do not provide reliably, and that the Darkboard evidence suggests is most critical precisely when the adversary is most sophisticated.

5.3 | Cyber Operations as Kriegspiel at Scale

Cyber conflict exhibits the Kriegspiel information structure with particular clarity [1]. An attacker's model of the defender's network is a metaposition: a belief state over possible network configurations consistent with observed responses to probes. The defender's model of the attacker's current position and capabilities is equally a metaposition. Both sides update incrementally based on ambiguous observations—network logs, intrusion detection alerts, and error messages, all of which can be falsified—and both have incentives to maximize information gain while minimizing information leakage, the exact probe-sequencing trade-off that Darkboard navigates in Kriegspiel. AI systems deployed in cyber defense that are designed primarily for high-throughput anomaly detection in benign environments will be structurally underprepared for adversaries who understand their detection logic and design intrusions to remain within its blind spots. The Kriegspiel-inspired design principles of Section 6—explicit belief-state representation, information-cost accounting, and robustness over point-estimate optimality—offer a more principled architectural foundation for AI in adversarial cyber environments.

5.4 | AI-Driven Strategies in International Relations

International relations constitute perhaps the most direct real-world analog of Kriegspiel's epistemic structure. States, like Kriegspiel players, observe only fragments of the global board: diplomatic signals are ambiguous, military deployments are partially concealed, and the intentions behind actions are systematically falsified. The challenge of interpreting a rival's nuclear posture, a covert cyber campaign, or an ambiguous troop concentration mirrors the metaposition problem—decision-makers must maintain a probability distribution over possible adversary configurations and act robustly across that distribution rather than committing to a single intelligence estimate. AI systems deployed in early-warning, treaty-verification, or crisis-management roles will face precisely this structure, and systems

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

trained exclusively on perfect-information benchmarks risk catastrophic overconfidence when adversaries deliberately manipulate the observable signal. The invisible chess family of benchmarks therefore offers foreign-policy AI designers a principled testbed that standard game-based evaluation cannot provide. A possibly appropriate response is not to withdraw AI decision-support from international affairs, but to adopt the centaur model: AI managing structured inference over observable signals, human judgment retaining authority over decisions where the belief state is adversarially contested and irreducibly uncertain [9].

5.5 | A Structured Mapping of Strategic Elements

Table 3 formalizes the correspondences developed in this section. Each row maps a specific Kriegspiel mechanism to its real-world counterpart and identifies a concrete contemporary case.

Two observations follow from the table. First, the formal properties of belief-state tracking under adversarial ambiguous feedback are shared by both the game and the real-world domain, not merely analogous by metaphor. A belief state maintained under adversarial observation, updated incrementally from ambiguous signals, and exploited by an adversary who models the update procedure is constitutive of any partially observable adversarial decision problem. Second, the table reveals a gap: no current invisible-chess variant fully models coalition dynamics—multiple agents with partially aligned objectives and private information—which is arguably the defining epistemic challenge of multinational conflict operations and alliance diplomacy. This gap points towards multi-agent POMDP extensions as the natural next step in the benchmark research agenda.

6 | Implications for AI Innovation and Governance

This section draws four clusters of implications from the preceding analysis: design principles for AI under uncertainty (Section 6), benchmark sovereignty, governance of high-stakes autonomous systems, including a concrete EU AI Act compliance scenario, and a proposed evaluation agenda.

The cyber-operations scenario introduced in Section 1 is one instantiation of a broader pattern: the following implications extend across AI-augmented conflict, digital sovereignty, and autonomous systems governance. In fact, the wargame-like "invisible chess" family of benchmarks converge on a set of implications for AI innovation policy that extend well beyond game-playing AI. We organize these around four themes: design principles for AI under uncertainty, benchmark sovereignty, governance of high-stakes autonomous systems, and regulatory applications.

6.1 | Three Design Principles for AI Under Uncertainty

6.1.1 | *Principle 1: Explicit Belief-State Representation*

AI systems in adversarial environments should maintain explicit, interpretable and auditable representations of their uncertainty—analogous to Darkboard's metaposition—rather than collapsing that uncertainty into a single confidence score [10]. A system that maintains a full belief state can detect when observations cause anomalous updates, a potential signature of adversarial manipulation, and alert human operators; a system that outputs only a point estimate cannot.

6.1.2 | *Principle 2: Information-Cost Accounting*

Decisions about whether to acquire additional information should be treated as first-class strategic choices with explicit cost accounting, following Darkboard's probe-sequencing model [8]. In adversarial environments, the act of probing reveals information to the adversary; AI systems should reason explicitly about the expected uncertainty reduction from different information-acquisition options, weighed against their operational and strategic costs.

6.1.3 | *Principle 3: Robustness Over Point-Estimate Optimality*

AI systems should maximize robust performance across the distribution of possible world states rather than optimal performance in the single most likely state [21]. Darkboard's

TABLE 3 | Kriegspiel–real-world analogy map.

Kriegspiel	Real-world	Example
Umpire feedback (illegal move/capture/check direction)	HUMINT: Partial, third-party-mediated signals	CIA source reporting on troop movements without unit ID [1]
Metaposition (belief state over hidden configurations)	Intelligence fusion/order-of-battle estimate	DIA all-source fusion; targeting packages under sensor denial [2]
False-structure injection (Case 1 in Section 4)	Deception operations/maskirovka	Soviet radio deception; operation bodyguard (1944) [1]
Probe-sequence baiting (Case 2 in Section 4)	Adversarial ISR triggering	Adversary spoofing radar to induce predictable scan reallocation [2]
Sensing budget allocation (RBC)	ISR collection management	Finite satellite passes tasked against prioritised targets [10]
Belief-state explosion under adversarial uncertainty	Cognitive overload/command-and-control (C2) degradation	Information overload in gulf war air operations [1]
Self-play training failure in imperfect-information domains	Training–deployment gap in real operations	AI systems trained on synthetic data failing under adversarial real-world conditions [32]

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

maximax sampling instantiates this principle: a move that performs well across many consistent configurations is preferred over one that is optimal for the assessed state but catastrophic if that assessment is wrong.

6.2 | Benchmarking Digital Sovereignty

National AI strategies—including the EU AI Act and the OECD AI Principles—frame digital sovereignty primarily in terms of data ownership, algorithmic independence, and supply-chain security [10, 33]. The Kriegspiel analysis suggests an additional dimension: benchmark sovereignty. The capabilities an organization develops are shaped by the benchmarks against which its systems are trained and evaluated. Organizations whose AI investments are dominated by perfect-information benchmarks are optimizing for a challenge that differs structurally from adversarial deployment [34, 35].

Benchmark sovereignty means investing in evaluation frameworks that reflect actual epistemic conditions. The OECD AI Principle on robustness and safety [10] implies that AI systems be tested in conditions representative of deployment; for adversarially deployed AI this requires imperfect-information benchmarks. The EU AI Act's conformity-assessment requirements for high-risk systems [33] similarly imply robustness testing under adversarial conditions that perfect-information benchmarks cannot satisfy. Imperfect-information chess variants—tractable, reproducible, and difficulty-scalable—constitute a natural candidate family for such certification environments.

To operationalize benchmark sovereignty, states should require adversarial, domain-specific benchmarking as part of procurement and conformity assessment for strategic AI systems. In practice, this means that any system intended for high-stakes security, defense, or critical infrastructure use should be evaluated against an approved suite of imperfect-information benchmarks—for example, Kriegspiel, Reconnaissance Blind Chess, or related partially observable environments—with results recorded in the technical documentation and reviewed during conformity assessment under Article 43 of the EU AI Act [33]. Where harmonized standards or common specifications do not yet exist, the benchmark suite can function as a common specification or procurement condition, allowing national authorities to require deception-robustness and uncertainty-handling tests before deployment.

6.3 | Governance of High-Stakes Autonomous Systems

The evidence found in invisible chess games adds a purely technical argument to established ethical and legal objections to full autonomy [2]: In adversarial environments, AI performance degrades systematically and non-linearly as adversaries learn to manipulate observation inputs, and this degradation is structural—a consequence of the computational geometry of imperfect-information decision-making—not a defect that better engineering will resolve. Decisions with irreversible high-stakes consequences should therefore not be delegated to systems whose performance degrades systematically under the conditions most likely to prevail in actual deployment.

The centaur model—AI managing tractable sub-problems while humans retain authority over decisions under adversarially maintained uncertainty—represents the technically optimal allocation of cognitive labor [9, 36]. This aligns with the EU AI Act's human-oversight requirements for high-risk AI [33] and with OECD AI Principle 1.4 on transparency and accountability [10].

More recent human–AI teaming research complicates the centaur analogy. Recent reviews of human–AI teaming show that hybrid teams do not reliably outperform humans or AI alone; instead, their success depends on clear role complementarity, calibrated trust, shared mental models, and task structure, whereas overreliance, poor coordination, and mismatched expectations can depress performance [37, 38]. Empirical work therefore suggests that centaur models are effective only under specific sociotechnical conditions, and that human involvement can fail when it adds override errors, slows adaptation, or reduces situation awareness rather than improving it. Accordingly, our argument for human–AI collaboration is conditional: the centaur model is valuable when the human contributes judgment, ethical authority, or contextual interpretation that the AI cannot provide, and when the division of labor is explicit and well calibrated.

6.3.1 | *An Illustrative Compliance Scenario*

Consider a hypothetical AI-based decision-support system (DSS) deployed by a national defense ministry to assist commanders in time-critical targeting decisions—a paradigmatic high-risk AI application under the EU AI Act [33].

Table 4 maps the system's functional requirements, as motivated by the invisible-chess analysis, onto specific EU AI Act provisions and OECD AI Principle obligations [10].

Under this mapping, a DSS that reports only a single confidence score (violating Principle 1) would fail to satisfy Art. 13 transparency requirements because a human overseer cannot meaningfully assess outputs whose uncertainty representation has been collapsed. A DSS whose performance against adversarially scripted opponents falls below the deception-robustness threshold proposed in Section 4.4 would similarly fail the Art. Nine conformity-assessment requirement that systems be tested under foreseeable adverse conditions. The invisible-chess benchmark suite thus provides not merely a research tool but a technically grounded, reproducible, and internationally comparable test environment that directly operationalizes existing regulatory obligations [10, 33].

6.4 | Applications of Partial Information Benchmarks

The value of these benchmarks extends well beyond defense [10, 33]. The following civilian domains illustrate the breadth of applicability.

6.4.1 | *Supply-Chain Resilience*

Supply-chain management under disruption is structurally analogous to Kriegspiel: a firm has partial knowledge of supplier

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 | Compliance scenario.

Requirement	EU AI act provision	OECD AI principle
Explicit belief-state representation and uncertainty logging (Principle 1)	Art. 13 (transparency): High-risk systems must provide information enabling human oversight of outputs and confidence levels	Principle 1.3 (transparency and explainability): AI actors should provide meaningful information about AI systems and their limitations
Human authority over adversarially contested decisions (centaur model)	Art. 14 (human oversight): High-risk systems must allow natural persons to oversee and intervene; override capability mandatory	Principle 1.4 (robustness, security, safety): AI systems should perform reliably and be overridable
Robustness testing against adversarial belief-state manipulation (Principle 3)	Art. 9 (risk management): Conformity assessment must include testing under foreseeable conditions of misuse, including adversarial inputs; Annex IV documentation must record test scenarios	Principle 1.5 (accountability): Organizations should be accountable for proper functioning and resilience under adversarial conditions
Imperfect-information benchmark suite as certification environment	Art. 40–43 (conformity assessment and notified bodies): Technical standards underpinning conformity assessment should be extended to cover adversarial imperfect-information test environments	Principle 1.3 and OECD recommendation 2.3: National AI strategies should invest in evaluation infrastructure representative of deployment conditions

capacity and competitor procurement activity; counterparties actively conceal vulnerabilities; decisions must be made under irreducible uncertainty about the hidden state. The three design principles above translate directly: maintain explicit uncertainty representations over possible supply-network configurations; model the information cost of probing suppliers; and select procurement strategies robust across plausible disruption scenarios. Imperfect-information game benchmarks can serve as tractable proxies for evaluating supply-chain AI robustness before deployment.

6.4.2 | *Competitive Business Strategy*

Firms making investment and R&D decisions under hidden-competitor strategies face a problem structurally analogous to RBC: they can sense only a limited region of the competitive landscape in each planning cycle and must act under uncertainty about the rest. Neural information-set weighting approaches [14] that concentrate sensing on strategically consequential regions translate into competitive-intelligence architectures that prioritize monitoring of the most relevant competitor activities. The mobility–visibility trade-off similarly captures the tension between committing resources to a market position and maintaining broad competitive awareness.

6.4.3 | *Regulatory Sandboxes and AI Certification*

AI certification frameworks currently rely predominantly on perfect-information benchmark test sets. The EU AI Act's conformity-assessment requirements [33] implicitly demand robustness evaluation representative of deployment, but existing certification infrastructure does not yet provide adversarial imperfect-information test environments. Incorporating performance on Kriegspiel-family benchmarks into AI certification frameworks would operationalize the OECD AI robustness principle [10] in a technically concrete, internationally comparable, and reproducible way—a practical step that follows directly from the analysis of this paper.

6.4.4 | *Clinical Decision-Making*

Diagnosis and treatment planning involve decision-making under partial observability and uncertainty, sharing deep structural similarity with Kriegspiel and RBC. The true disease state remains hidden: the clinician relies on noisy, incomplete, and occasionally misleading signals from symptoms, biomarkers, and imaging. Each diagnostic test consumes a limited sensing budget—analogous to RBC's constrained 3×3 sensing window—while carrying costs, risks, and the potential for false signals. Moreover, pathogens and tumors evolve in response to treatment, actively corrupting the clinician's belief state in a manner reminiscent of adversarial metaposition injection. Just as Kriegspiel was developed to train officers in decision-making under the fog of war, medical education has long employed wargame-style simulations for high-stakes clinical reasoning [39]. Imperfect-information benchmarks therefore offer a principled, controllable environment for evaluating clinical AI before deployment in safety-critical settings.

6.5 | A Proposed Evaluation Agenda

The three design principles and governance arguments above remain most actionable if grounded in concrete, reproducible evaluation practices. We propose a minimal agenda organized around three components.

6.5.1 | *Standardized Benchmark Suite*

A shared invisible-chess evaluation suite should include at minimum one variant from each information-restriction regime: Kriegspiel (umpire-mediated, reactive feedback), RBC (explicit sensing budget), and Dark Chess or Fog of War (passive visibility only). Performance should be reported on three metrics: belief-entropy reduction rate (how efficiently the agent narrows

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the belief state per move), sensing efficiency (ratio of strategically relevant information gained to sensing actions consumed, applicable to RBC), and deception robustness score (win-rate degradation against scripted adversaries that maximize opponent belief entropy, relative to a uniform-play baseline). These metrics are computable from game logs without access to ground-truth hidden state, making them suitable for external certification audits.

6.5.2 | *Failure-Mode Thresholds for Autonomous Systems*

Governance claims about the unsuitability of autonomous systems for high-stakes decisions become operationally concrete when tied to measurable failure modes. As a starting point, we propose that an AI system be classified as unsuitable for autonomous weapons if its deception-robustness score falls below a defined threshold under standardized adversarial play—analogous to safety-critical thresholds in aviation or medical AI certification. The invisible-chess suite provides a tractable environment in which such thresholds can be empirically calibrated before systems are deployed in operational contexts.

6.5.3 | *Integration Into National AI Evaluation Frameworks*

National AI strategies and the EU AI Act's conformity-assessment process [33] should require imperfect-information benchmark testing for AI systems designated for adversarial or high-risk deployment. Kriegspiel-family variants satisfy the key properties needed for regulatory use: they are formally specified, publicly reproducible, difficulty-scalable (by adjusting time controls and opponent strength), and internationally comparable without requiring access to classified operational data [10].

7 | Conclusion

This article has made three contributions. First, at the taxonomic level, it has proposed a structured taxonomy of imperfect-information chess variants—Kriegspiel, Dark Chess, Fog of War, and RBC—organized by the type and degree of information asymmetry they impose, as a richer and more faithful benchmark family for strategic AI than perfect-information chess. Second, at the empirical level, the Darkboard case study has grounded three propositions in a decade-long competitive record: that AI can achieve strong performance when the belief state is locally tractable; that performance degrades systematically when belief-state complexity is adversarially maintained; and that skilled human players retain a meaningful advantage precisely in this second regime. Third, at the governance level, these findings carry structural implications for the design and evaluation of AI systems operating in adversarial real-world environments—specifically, against delegating irreversible high-stakes decisions to autonomous systems where an opponent can model and exploit the agent's inference procedure, and in favor of centaur models of human-AI collaboration.

This paper has argued that the standard chess analogy for AI-augmented strategy—powerful, widely cited, and genuinely instructive in limited respects—rests on a structural misrepresentation of the epistemic conditions that define real strategic environments [40]. Chess is a game of perfect information; warfare, cyber conflict, and international relations are not. The analogy illuminates AI's capacity for deep combinatorial reasoning while concealing its vulnerability to adversarial uncertainty which constitutes the defining challenge of strategic competition. Imperfect-information chess variants—and Kriegspiel in particular—provide a richer and more faithful analog, one that preserves the combinatorial depth and adversarial structure of chess while incorporating the hidden state, deceptive observation, and belief-state reasoning that characterize real conflict. At the same time, this analogy remains incomplete: real conflict also involves coalition dynamics, civilian populations, physical terrain, normative constraints, and psychological factors that no chess variant can capture. For that reason, the policy implications of this paper should be read as conditional and mechanism-based rather than deductive or categorical.

The argument has been developed across three registers. Theoretically, we have shown that the transition from perfect-information to imperfect-information games involves not merely a quantitative increase in computational difficulty but a qualitative change in the nature of the AI challenge: from evaluating a known game tree to reasoning over a distribution of possible trees, maintaining a belief state under adversarial manipulation, and selecting actions that are robust across epistemic uncertainty rather than optimal for a single known state [7]. Empirically, the Darkboard case study has grounded these theoretical claims in the concrete record of a system built to address precisely these challenges: demonstrating where principled belief-state AI achieves strong and, in some cases, optimal performance, and documenting the specific failure modes—belief-state explosion, adversarially engineered meta-position corruption, and the irreducible human edge in opponent modeling—that reveal the structural limits of AI under persistent adversarial uncertainty [8]. Practically, the analysis has derived three design principles for AI decision-support in strategic environments, a technically grounded argument for the centaur model of human-AI command, a novel dimension of digital sovereignty centered on benchmark choice, and a purely technical argument—independent of ethical and legal considerations—against the delegation of lethal engagement decisions to autonomous systems in adversarially contested environments [2]. However, these conclusions support stricter testing, mandatory human oversight, and constrained delegation rather than a direct deductive case for categorical prohibition.

Several limitations of the present analysis should be acknowledged. The Darkboard evidence, although uniquely authoritative for a paper of this kind, derives from a single system developed within a specific technical paradigm; stronger future Kriegspiel engines using model-based reinforcement learning or large-scale neural self-play may achieve higher performance, though we have argued that the structural obstacles to Kriegspiel AI dominance are unlikely to be fully overcome by scaling existing techniques [41]. The transfer from Kriegspiel to real strategic environments is analogical, not deductive: real-world conflicts involve dimensions of uncertainty—coalition dynamics, civilian populations, physical and psychological human factors, normative constraints—that no chess variant can capture. The paper's

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

claims about AI limits in strategic environments should be read as hypotheses motivated by game AI evidence, requiring further validation through dedicated empirical studies of AI systems in operational or high-fidelity simulated military contexts. Accordingly, the policy recommendations advanced here are best understood as provisional guidance for settings where hidden state, deception, and adversarial inference are central, not as universal prescriptions for all forms of conflict.

Notwithstanding these limitations, the core contribution stands. The chess-to-warfare analogy is not wrong; it is incomplete. Completing it requires moving from the perfect-information boardroom of Kasparov and Deep Blue to the fog-of-war laboratory of Kriegspiel and Darkboard. In that laboratory, AI systems encounter the epistemic conditions of real strategic competition—hidden state, adversarial deception, irreducible uncertainty—and their performance reveals both genuine capabilities and genuine limits. Understanding those limits is not counsel of AI skepticism; it is the precondition for building AI systems that are genuinely robust in the environments that matter most, and for designing the human-AI partnerships that the strategic realities of the 21st century will require. Accordingly, we foreground the defense and international-relations argument, where the link between hidden state, deception, and adversarial uncertainty is strongest, and treat the other analogies as illustrative rather than central.

The broader lesson for AI innovation policy is that benchmark choice is a strategic decision. Nations and organizations that evaluate their strategic AI systems against perfect-information benchmarks are optimizing for a world that does not exist in conflict settings. Investing in imperfect-information research—such as Kriegspiel, reconnaissance blind chess, multi-agent partially observable environments, and the theoretical frameworks that unify them—is not a diversion from applied AI development; it is the foundation on which strategically robust AI can be built.

Author Contributions

Paolo Ciancarini: conceptualization, investigation, writing – original draft.

Acknowledgements

The author thanks Andrea Bolognesi, Gian Piero Favini, and Andrea Gasparro, who contributed to the Darkboard project over several years.

Funding

The author has nothing to report.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. L. Freedman, *The Future of War: A History* (PublicAffairs, 2017): ISBN 9781541742468.

2. M. L. Cook and C. Brose, "The Kill Chain: Defending America in the Future of High-Tech Warfare," *Naval War College Review* 73, no. 4 (2020): 11, https://digital-commons.usnwc.edu/nwc-review/vol73/iss4/11.

3. F.-H. Hsu, *Behind Deep Blue* (Princeton University Press, 2002): ISBN 9780691118185.

4. T. Romstad, M. Costalba, J. Kiiski, and G. Linscott, and the Stockfish Community, Stockfish: A Strong Open-Source Chess Engine (2008), https://stockfishchess.org.

5. D. Silver, T. Hubert, J. Schrittwieser, et al., "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go Through Self-Play," *Science* 362, no. 6419 (2018): 1140–1144, https://doi.org/10.1126/science.aar6404.

6. J. F. Dunnigan, *Zones of Control: Perspectives on Wargaming* (MIT Press, 2016).

7. J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, 1944).

8. P. Ciancarini and G. P. Favini, "Monte Carlo Tree Search in Kriegspiel," *Artificial Intelligence* 174, no. 11 (2010): 670–684, https://doi.org/10.1016/j.artint.2010.04.017.

9. G. Kasparov and M. Greengard, *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins* (PublicAffairs, 2017): ISBN 9781610397650.

10. OECD, *State of Implementation of the OECD AI Principles: Insights From National AI Policies*. Technical report (Organisation for Economic Co-operation and Development, 2021), https://oecd.ai/en/dashboards/overview.

11. C. E. Shannon, "Programming a Computer for Playing Chess," *Philosophical Magazine* 41, no. 314 (1950): 256–275, https://doi.org/10.1080/14786445008521796.

12. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *Poster of the 3rd International Conference on Learning Representations (ICLR)*, (2015), https://doi.org/10.48550/arXiv.1412.6572.

13. B. Hu Zhang and T. Sandholm, "General Search Techniques Without Common Knowledge for Imperfect-Information Games, and Application to Superhuman Fog of War Chess," in *Poster of the 14th International Conference on Learning Representations (ICLR)*, (2026), https://doi.org/10.48550/arXiv.2506.01242.

14. T. Bertram, J. Furnkranz, and M. Muller, "Neural Network-Based Information Set Weighting for Playing Reconnaissance Blind Chess," *IEEE Transactions on Games* 16, no. 4 (2024): 960–970, https://doi.org/10.1109/TG.2024.3425803.

15. G. Clark, "Deep Synoptic Monte-Carlo Planning in Reconnaissance Blind Chess," *Advances in Neural Information Processing Systems* 34 (2021): 4106–4119, https://proceedings.neurips.cc/paper_files/paper/2021/file/215a71a12769b056c3c32e7299f1c5ed-Paper.pdf.

16. M. Augier, N. Dew, T. Knudsen, and N. Stieglitz, "Organizational Persistence in the Use of War Gaming and Scenario Planning," *Long Range Planning* 51, no. 4 (2018): 511–525, https://doi.org/10.1016/j.lrp.2017.12.005.

17. A. Matros, "Lloyd Shapley and Chess With Imperfect Information," *Games and Economic Behavior* 108 (2018): 600–613, https://doi.org/10.1016/j.geb.2017.12.003.

18. M. Sakuta and H. Iida, "Solving Kriegspiel-Like Problems: Exploiting a Transposition Table," *ICGA Journal* 23, no. 4 (2000): 218–229, https://doi.org/10.3233/ICG-2000-23403.

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19. T. Mellemseter, *Reasoning Under Uncertainty in Darkchess* Master's Thesis (Norwegian University of Science and Technology (NTNU), 2017), https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2448943.

20. R. W. Gardner, "The Machine Reconnaissance Blind Chess Tournament of Neurips 2022," in *Proceedings of NEURIPS* (PMLR, 2023), 119–132, https://proceedings.mlr.press/v220/gardner23a.

21. P. Ciancarini and G. Piero Favini, "Representing Kriegspiel States With Metapositions," in *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI)*, (2007), 2450–2455, https://www.ijcai.org/Proceedings/07/Papers/394.pdf.

22. A. Bolognesi and P. Ciancarini, "Computer Programming of Kriegspiel Endings: The Case of KR vs K," in *Advances in Computer Games: Many Games, Many Challenges* H. Jaap Van Den Herik, H. Iida, and E. A. Heinz, eds. (Springer, 2004), 325–341, https://doi.org/10.1007/978-0-387-35706-5_21.

23. P. Ciancarini and G. Piero Favini, "Playing the Perfect Kriegspiel Endgame," *Theoretical Computer Science* 411, no. 40–42 (2010): 3563–3577, https://doi.org/10.1016/j.tcs.2010.05.019.

24. G. P. Favini, The Dark Side of the Board: Advances in Chess Kriegspiel. PhD thesis, (2010): Department of Computer Science, http://www.cs.unibo.it/pub/TR/UBLCS/2010/2010-06.pdf.TechnicalReportUBLCS-2010-06.

25. L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and Acting in Partially Observable Stochastic Domains," *Artificial Intelligence* 101, no. 1–2 (1998): 99–134, https://doi.org/10.1016/s0004-3702(98)00023-x.

26. S. Russell and J. Wolfe, "Efficient Belief-State AND–OR Search, With Application to Kriegspiel," in *Proceedings of IJCAI*, (2005), 278–285, https://www.ijcai.org/Proceedings/05/Papers/0929.pdf.

27. C. H. Papadimitriou and J. N. Tsitsiklis, "The Complexity of Markov Decision Processes," *Mathematics of Operations Research* 12, no. 3 (1987): 441–450, https://doi.org/10.1287/moor.12.3.441.

28. M. Mundhenk, J. Goldsmith, C. Lusena, and E. Allender, "Complexity of Finite-Horizon Markov Decision Process Problems," *Journal of the ACM* 47, no. 4 (2000): 681–720, https://doi.org/10.1145/347476.347480.

29. D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The Complexity of Decentralized Control of Markov Decision Processes," *Mathematics of Operations Research* 27, no. 4 (2002): 819–840, https://doi.org/10.1287/moor.27.4.819.297.

30. J. H. Reif, "The Complexity of Two-Player Games of Incomplete Information," *Journal of Computer and System Sciences* 29, no. 2 (1984): 274–301, https://doi.org/10.1016/0022-0000(84)90034-5.

31. F. P. B. Osinga, *Science, Strategy and War: The Strategic Theory of John Boyd*. volume 18 of Strategy and History (Routledge, 2007): ISBN 9780415459525, https://doi.org/10.4324/9780203088869.

32. R. Sheatsley, N. Papernot, M. J. Weisman, G. Verma, and P. McDaniel, "Adversarial Examples for Network Intrusion Detection Systems," *Journal of Computer Security* 30, no. 5 (2022): 727–752, https://doi.org/10.3233/JCS-210094.

33. European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)," *Official Journal of the European Union, L 2024/1689* (2024), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689.

34. Office of the Director of National Intelligence, "The AIM Initiative: A Strategy for Augmenting Intelligence Using Machines," *Technical report, Office of the Director of National Intelligence* (2020), https://www.dni.gov/files/ODNI/documents/AIM-Strategy.pdf.

35. A. Goldfarb and J. R. Lindsay, "Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War," *International Security* 46, no. 3 (2022): 7–50, https://doi.org/10.1162/isec_a_00425.

36. B. Jensen and Y. Atalan, *The Pentagon's AI Problem Isn'T Algorithms, It's Evaluation* (Center for Strategic and International Studies (CSIS), 2025), https://www.csis.org/analysis/pentagons-ai-problem-isnt-algorithms-its-evaluation.

37. C. Gonzalez, K. Donahue, D. G. Goldstein, et al., "Toward a Science of Human–AI Teaming for Decision Making: A Complementarity Framework," *PNAS Nexus* 5, no. 3 (2026): pgag030, https://doi.org/10.1093/pnasnexus/pgag030.

38. National Academies of Sciences, Engineering, and Medicine, *Human-AI Teaming: State-of-the-Art and Research Needs*. Technical report (National Academies Press, 2022), https://www.nationalacademies.org/read/26355.

39. T. L. Hedrick and J. S. Young, "The Use of 'War Games' to Enhance High-Risk Clinical Decision-Making in Students and Residents," *American Journal of Surgery* 195, no. 6 (2008): 843–849, https://doi.org/10.1016/j.amjsurg.2007.06.032.

40. D. Rasskin-Gutman, *Chess Metaphors: Artificial Intelligence and the Human Mind* (MIT Press, 2012).

41. J. Schrittwieser, I. Antonoglou, T. Hubert, et al., "Go, Chess and Shogi by Planning With a Learned Model," *Nature* 588 (2020): 604–609, https://doi.org/10.1038/s41586-020-03051-4.

30673941, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aiv2.70012 by Statens Beredning, Wiley Online Library on [01/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License