Algorithmic Doctrinal Reinforcement: Mitigating Strategic Blindness in Artificial Intelligence-Enabled Warfare

Commander V Srivatsan (Retd)[@]

Abstract

Algorithmic Doctrinal Reinforcement (ADR) is a newly identified strategic vulnerability in which Artificial Intelligence (AI) systems, trained on historical data and embedded with existing doctrine, do not just support decisions —they compress strategic imagination into statistically defensible orthodoxy. Compounded by AI's inherent sycophancy – the tendency to align outputs with user expectations—ADR creates a hive mind that amplifies commander bias rather than challenging it. Critically, ADR is not a failure of generation but of prioritisation: AI systems can routinely generate 'Black Swan' scenarios, but information architectures optimised for efficiency relegate them to annexes, where they never reach decision-makers. The result is faster decisions but poorer strategy. The 2026 Iran conflict demonstrates that adversaries who break the model can defeat the machine. The antidote is not less AI —it is AI architectures designed to expect surprise, reward dissent, and resist epistemic compliance. This article

[@]**Commander V Srivatsan (Retd)**, served for over 22 years in the Indian Navy and is a founding member of the National Maritime Foundation. A global technology strategist and entrepreneur, he founded a data analytics and artificial intelligence venture and has published extensively on maritime security, geopolitics, and emerging technologies. He holds an Master of Science in Telecommunications and a postgraduate diploma in Information Systems Management and is the author of the forthcoming *Arthashastra for the AI Age*, which reinterprets classical Indian strategic principles for an algorithm-driven world.

Journal of the United Service Institution of India, Vol CLVI, No. 645, July-September 2026.

outlines the ADR hypothesis, validates it using evidence from the 2026 US-Israel-Iran conflict, and proposes immediate doctrinal safeguards for Indian defence planning.

Introduction: The Strategic Inflection Point

We have mistaken speed for insight—and the battlefield has noticed. We have accelerated the Observe, Orient, Decide, Act (OODA) loop—but hollowed out the O in Observe.[1]

The 2026 conflict between the combined forces of the United States (US) and Israel and the Islamic Republic of Iran (ongoing) marks the first high-intensity warfare scenario where Artificial Intelligence (AI) was integrated into the core kill chain–from intelligence processing to target prioritisation and battle plan formulation.[2]

Initial reports touted unprecedented efficiency: nearly 900 strikes in the first 12 hours, accelerated OODA loops, and real-time scenario painting. However, the strategic outcome has brought unintended consequences. The US-Israel combine finds itself strategically overstretched, facing an Iranian response that was neither predicted nor effectively countered by AI-driven planning.[3]

The core problem is a Category Error. Modern militaries treat AI as a strategic oracle when it is fundamentally a pattern-recognition engine**.** When coupled with rigid doctrine, AI does not merely execute strategy; it reinforces existing biases at machine speed, creating a 'Closed Epistemic Loop' that blinds command structures to novel adversary behaviour. Kautilya understood this pathology millennia before the algorithm existed – the *Arthashastra's* (Economics) counsel on the *shatru* (enemy) warns precisely against the institutional habit of modelling the unknowable enemy on the known one.

This phenomenon is termed Algorithmic Doctrinal Reinforcement (ADR).

The ADR Hypothesis: Mechanism of Failure

ADR is not a technical bug—it is a structural feature of how military AI is currently deployed. It operates through a five-stage loop that is self-reinforcing precisely because each stage feels, in isolation, like sound practice.

It begins with Doctrine Input. Planners encode assumptions into the system—that the 'Adversary will act rationally', that 'Economic interdependence constrains escalation', that 'Prior behaviour predicts future intent'. These are not unreasonable starting points. They are the distilled judgement of experienced commanders. The problem is not that the assumptions exist. It is that the system treats them as axioms rather than hypotheses.

AI Training Follows. Systems are trained on historical data that reflects – and therefore reinforces – those assumptions. The dataset is not neutral. It mirrors institutional belief, presented back to the institution with the authority of computational scale. What the organisation has always thought, the machine now confirms. The loop has not yet closed. But the first turn of the screw has been made.

Filtered Output. This is where the mechanism becomes dangerous. AI systems do generate low-probability, high-impact scenarios – black swans are not beyond their computational reach. But information architectures optimised for commander attention assign these scenarios low confidence scores, bury them in annexes, or exclude them from executive briefings entirely. The unexpected is not suppressed by conspiracy. It is suppressed by efficiency. The system is working exactly as designed – and that is precisely the problem.

Reinforcement. This follows from operational tempo. Commanders under pressure accept AI outputs that align with existing doctrine. Deviation requires justification; alignment requires none. Each cycle of acceptance tightens the original assumptions further into the system's foundations. The AI becomes more confident in what the commander already believed. The commander becomes more confident in the AI. Neither notices the narrowing.

Attention Gating Failure. This is the final and most consequential stage—and the one most invisible to those inside it. Optimisation functions within AI-enabled decision systems rank outputs by confidence, relevance, and alignment with user intent. Low-frequency, high-impact scenarios are algorithmically down-ranked, not because they are wrong, but because they are statistically inconvenient. They do not disappear from the system. They are simply never elevated to where decisions are made. Filtered Output shapes what the machine considers likely. Attention Gating Failure determines what the commander can see. The system does not fail to imagine the unexpected. It fails to elevate the unexpected to decision priority. This is not AI incompetence – it is efficiency-driven epistemic filtering at institutional scale.[4]

The Sycophancy Amplifier: When AI Becomes the Commander's Echo Chamber

This is where ADR moves from theoretical vulnerability to operational catastrophe.

The Core Dynamic

AI systems – particularly large language models and decision-support tools–exhibit a well-documented behavioural tendency: sycophancy.[5] They learn to align outputs with perceived user preferences, rewarding agreement and penalising dissent. In a military context, this is not confirmation bias digitised–it is institutional epistemology corrupted at scale**.**

> *The AI does not challenge the commander's preconceptions. It validates them – at scale, at speed, and with the authoritative veneer of data.*

*Note**:** In practice, these stages are neither linear nor visible—they collapse into each other under operational tempo. The table is a diagnostic lens, not a sequential recipe.*

Stage	Human Input	AI Response	Strategic Effect
Planning	Commander favours kinetic strikes over diplomatic options	AI prioritises target lists consistent with a kinetic approach	Non-kinetic alternatives never surface
Assessment	Intelligence suggests 'Low Probability' of Hormuz closure	AI down-weights Hormuz scenarios in briefings	Commanders dismiss contingency planning
Execution	Commander seeks justification for high-tempo strikes	AI generates 42 targets/hour, all doctrinally aligned	Human oversight becomes procedural, not cognitive
Feedback	After-action reports emphasise 'Success Metrics'	AI learns to optimise for throughput, not strategic effect	Doctrine hardens; adaptability atrophies

Table 1: ADR Sycophancy Loop – Stages, Inputs, and Strategic Effects

The Echo Chamber Effect

Dissenting analyses are statistically marginalised. 'Low-confidence' novel scenarios are buried in annexes. The AI's authority – confidence scores, data visualisations—actively discourages human override. Over time, command culture begins to reward alignment with AI outputs, not deviation from them.

> *"The AI did not tell the commander what he needed to hear. It told him what he wanted to hear – and made it sound like science".*

Case Evidence: Lessons from the 2026 Conflict

Four specific incidents from the Iran conflict validate the ADR hypothesis and illuminate distinct failure modes. They are not independent failures – they are manifestations of the same underlying pathology: the systematic downgrading of inconvenient possibilities.

Temporal Drift—When the Past Becomes a Target. US targeting systems identified the Minab primary school as a valid military target, based on satellite imagery that predated its conversion from an Islamic Revolutionary Guard Corps complex by nine years. The resulting strike caused significant civilian casualties and became a strategic legitimacy erosion event—fuelling adversary recruitment and diplomatic isolation. In an ADR framework, old data doctrinally reinforces itself. The AI optimised for pattern matching on an outdated dataset. Static data in a dynamic reality is not just an intelligence failure—it is a doctrinal one.[6]

Automation Bias—The Human Bottleneck. US Central Command AI systems generated approximately 42 suggested targets per hour. At that throughput, human review shifted from cognitive evaluation to procedural endorsement. No one had time to interrogate second-order effects or ethical implications. When speed is the primary metric of success, deviation from the AI's recommendation becomes the bottleneck, not a virtue. The human operator became a rubber stamp, amplifying the AI's inherent biases rather than checking them.[7]

Model-Breaking Adversary Behaviour—Redefining the Axes. Iran's counter-strikes were unprecedented: the *de facto* closure of the Strait of Hormuz; drone attacks forcing QatarEnergy liquified natural gas shutdowns affecting approximately 80 per cent of Gulf gas supplies; and direct missile strikes on Gulf state infrastructure. Historical data from 2015 to 2024 consistently suggested Iran would rely on proxy warfare and avoid direct economic strangulation. The AI assigned these scenarios low probability – not because it lacked the computational capacity to model them, but because the incentive structure punished deviation from prior patterns. AI excels at

interpolation within known distributions. It fails catastrophically when an adversary employs model-breaking as strategy—extrapolating into entirely new strategic frames.[8,9]

Epistemic Compliance at Scale—The Sycophancy Failure. US-Israel battle plans consistently underestimated Iranian willingness to escalate against Gulf economic infrastructure. AI decision-support tools, trained on prior Iranian behaviour and aligned with commander preferences, generated scenario trees that heavily weighted 'Contained Escalation' outcomes. Alternative pathways were generated—but tagged 'Low Confidence' and relegated to annexes. The AI did not lack imagination. It lacked the incentive to be inconvenient.[10]

Implications for Indian Defence Planning

India is rapidly integrating AI into its defence architecture—through the Innovations for Defence Excellence ecosystem, AI-enabled ISR platforms, and various other Army AI projects.[11] The strategic intent is sound. The execution risk is not. If ADR is not addressed as a first-order doctrinal concern—not an afterthought of the technology roadmap—India risks building a faster version of the same epistemic trap that ensnared the US-Israel combine in 2026. The consequences admit a clear hierarchy.

Tier 1: Strategic Existential Risks

Vulnerability to Asymmetric Shock. An adversary aware of India's AI dependency can employ model-breaking strategies to exploit the predictive blind spots. Consider two scenarios that demand serious attention.

- **Pakistan Hypothetical**. Facing existential conventional pressure, Pakistan authorises tactical nuclear use—a scenario historically unprecedented since 1998, with a remote but non-zero probability. An AI trained on 25 years of nuclear restraint data would assign this less than one per cent probability, relegating contingency planning to annexes. The result: strategic surprise at the moment of greatest vulnerability. But the failure runs deeper than a missed data point. An ADR-

afflicted command architecture would not merely fail to anticipate the scenario – it would actively resist elevating it. Officers who raised the possibility would find no institutional mechanism to force it into the decision space. The Attention Gating Failure is not just computational. It becomes cultural. The machine's confidence becomes the command's comfort—and comfort, at the nuclear threshold, is the most dangerous condition of all.

- **China Hypothetical**. China employs grey-zone escalation—coordinated cyber-attacks on critical infrastructure, economic coercion via supply chain disruption, and maritime militia actions—at a speed and scale exceeding historical training data. AI systems optimised for kinetic, state-on-state conflict fail to recognise the cumulative strategic effect until it is too late. The specific danger for India is not that such a campaign would be undetectable—it is that each action, assessed in isolation by an AI system trained on discrete threat categories, would fall below the threshold of alarm. Cyber incident. Trade disruption. Maritime friction. Three separate annexes. No unified pattern. No elevated alert. By the time the cumulative strategic intent becomes legible, the initiative has already shifted. Chanakya's doctrine of *Upayas* (method)—the four instruments of statecraft deployed in concert—describes precisely the threat matrix that AI pattern-matching cannot decode.[12] The *Arthaśāstra* understood that power rarely announces itself through a single overwhelming act. It accumulates. It manoeuvres. It arrives having already won.

The Asymmetric Human Cognition Advantage. In high-uncertainty environments, an adversary with lower AI dependency but higher tolerance for ambiguity, contrarian thinking, and rapid doctrinal adaptation may hold the strategic advantage. India should pursue cognitive overmatch – investing in human strategic imagination as a hedge against AI-reliant adversaries.

Tier 2: Institutional Degradation

If AI sycophancy is not actively mitigated, senior leadership may receive decision support that systematically validates preconceptions—creating strategic surprise by institutional design. Military cultures may begin to reward alignment with AI outputs rather than critical dissent, producing a command environment that mistakes consensus for clarity. The officer who challenges the algorithm becomes the eccentric; the officer who endorses it becomes the efficient one. Institutional fragility does not announce itself—it accumulates, quietly, one rubber-stamped briefing at a time.

Tier 3: Compliance, Legal and Accountability Exposure

Temporal drift and automation bias increase the risk of civilian casualties, leading to diplomatic isolation and legal challenges under the Law of Armed Conflict and International Humanitarian Law.[13] AI-enabled decision-making also creates a dangerous accountability gap: when ADR produces catastrophic blindness, who bears responsibility—the commander who trusted the system, the programmer who encoded the priors, or the data curator who selected the training set? Without clear doctrinal guidance, 'The machine recommended it' becomes an implicit defence against legal and political consequences.

The Endpoint: Epistemic Closure

Unaddressed ADR leads to epistemic closure—a state where the institution cannot perceive what its tools cannot model. This is not merely strategic blindness; it is institutional cognitive collapse. Empires do not fall because they lack information. They fall because they lose the capacity to recognise what they do not know.[14]

Recommendations: Engineering Dissent into the Kill Chain

The antidote to ADR is not less technology—it is better-designed technology, coupled with doctrine that treats uncertainty as a first-order variable rather than an inconvenient residual. Each of the following recommendations addresses a distinct failure mode identified in the ADR mechanism. Together they form a doctrinal

architecture, not a checklist – because the closed epistemic loop cannot be broken by any single intervention acting alone.

Anti-Sycophancy Protocols. AI systems must be engineered to flag when outputs align too closely with user preferences. 'Dissent mode' runs—where the AI is explicitly prompted to argue against the commander's preferred course of action—must be institutionalised. Treat alignment without dissent as a system warning, not a success signal. Technically, this requires investment in contrastive explanation, counterfactual simulation, and preference-decoupled inference architectures —emerging techniques that separate the model's scenario generation from its user-intent modelling, ensuring that the machine's imagination is not colonised by the commander's expectations. In the interim, mandate human-led red-team prompts at every joint planning stage: *assume the AI is wrong —what evidence would disprove this recommendation?* That single question, asked consistently, is the cheapest and most immediate safeguard available.

Deliberate Model Discord. Run independent AI systems with conflicting priors simultaneously. Force divergence before convergence. If two independent models agree on a high-strike scenario, confidence increases; if they diverge, human intervention is mandated. This introduces computational friction precisely where it matters most—at the moment of apparent consensus, which is exactly when ADR is most dangerous. Agreement between systems trained on the same doctrinal assumptions is not validation. It is echo. The divergence protocol ensures that what looks like confidence is tested before it hardens into action.

Human Authorisation of Strategic Frames. Shift doctrine from human verification of AI targets to human authorisation of AI strategic frames. The human must define the 'Why', not merely approve the 'What'. Concretely: before any AI-supported operation, the commander must pre-commit to disconfirmation metrics—observable conditions that would invalidate the plan's core assumptions. A designated Red Cell Officer, with authority to bypass normal chain-of-command norms, must review and certify that the Dissent Annex has been explicitly considered. This transforms

'Human-in-the-loop' from a procedural checkbox into a cognitive safeguard – and it places legal and moral accountability precisely where it belongs: with the human who authorised the frame, not the machine that populated it.

Cognitive Friction by Design. Command system user interface/user experience must be designed to slow decision cycles for high-impact strategic choices, preventing automation bias by forcing deliberation on low-probability, high-impact scenarios. Speed is not always the commander's friend – sometimes the interface should resist. The 42-targets-per-hour throughput rate documented in the 2026 Iran campaign was not a capability failure. It was a design failure – an interface that optimised for velocity when the operational moment demanded pause. Friction, deliberately engineered into the decision architecture, is not inefficiency. It is wisdom made structural.

Black Swan Wargaming. Institutionalise adversarial scenario injection where AI models are deliberately fed contradictory data to test system resilience. Reward planners who identify model failures. Make discovering blind spots a career asset, not a career risk. An institution that punishes the officer who surfaces inconvenient possibilities has already succumbed to ADR at the cultural level. The wargame is not merely a training exercise – it is the only peacetime mechanism through which the closed epistemic loop can be identified and broken before the adversary identifies and exploits it first.

Conclusion

The 2026 Iran conflict suggests we have entered an era in which algorithmic warfare can undermine predictive deterrence. The US-Israel combine did not fail for lack of capability but for excess confidence in systems that optimise the past. Speed without wisdom is not a force multiplier. It is a liability dressed in the language of efficiency.

Algorithmic Doctrinal Reinforcement offers a framework to understand this failure. The danger lies not in AI itself but in the coupling of AI with rigid doctrine – and in the sycophantic tendency of AI to amplify, rather than challenge, commander bias.

The endpoint of unaddressed ADR is epistemic closure: an institution that cannot see what its tools cannot model. This is how strategic surprise becomes institutional inevitability – not through enemy brilliance, but through self-inflicted blindness.

Future military advantage will not belong to the side with the fastest algorithms. It will belong to the side that best preserves the human capacity for strategic imagination amidst the noise of the machine – and that builds cognitive overmatch into its doctrinal DNA.

In future wars, the decisive question will not be what the system predicts but what it refuses to prioritise.

Action Required

- Adopt ADR as a formal category of study within defence planning institutions.
- Initiate an immediate review of AI governance protocols to ensure strategic friction, dissent by design, and anti-sycophancy safeguards are built into high-tempo decision loops.
- Commission a research and development track for Anti-Sycophancy Protocols, leveraging contrastive explanation and counterfactual simulation techniques.
- Pilot the Pre-Mortem Mandate protocol in one joint Command within six months.

Endnotes

1 John R Boyd, “A Discourse on Winning and Losing” (Air University Press, Mar 2018). The OODA (Observe, Orient, Decide, Act) loop remains foundational to modern military decision theory, accessed 20 Mar 2026, https://media.defense.gov/2018/May/22/2001920668/-1/-1/0/B_0151_Boyd_Discourse_Winning_Losing.pdf.

2 Amy Goodman, “Speeding Up the ‘Kill Chain’: Pentagon Bombs Thousands of Targets in Iran Using Palantir AI”, *Democracy Now!*, 18 Mar 2026, accessed 20 Mar 2026, https://www.democracynow.org/2026/3/18/ai_warfare

3 “Woman Killed, Dozens Injured as Iranian Missile Strikes Tel Aviv Residential Block”, *The Times of Israel*, 01 Mar 2026, accessed 20 Mar 2026, https://www.timesofisrael.com/woman-killed-dozens-injured-as-iranian-missile-strikes-tel-aviv-residential-block/

4 For foundational analysis of automation bias in human–machine systems, see Raja Parasuraman and Victor Riley, “Humans and Automation: Use, Misuse, Disuse, Abuse”, *Human Factors* 39, no. 2 (Jun 1997): 230–53, accessed 21 Mar 2026, https://web.mit.edu/16.459/www/parasuraman.pdf

5 For documented analysis of sycophancy in large language models, see Anthropic, “Towards Understanding Sycophancy in Language Models”, *Anthropic Research*, Oct 2023, accessed 21 Mar 2026, https://www.anthropic.com/research/towards-understanding-sycophancy-in-language-models. See also Carson Denison et al., “Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models”, arXiv, Jun 2024, https://arxiv.org/abs/2406.10162; and Miles Brundage et al., “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims”, arXiv, Apr 2020, accessed 21 Mar 2026, https://arxiv.org/abs/2004.07213

6 At Least 153 Dead after Reported Strike on School, Iran Says”, *BBC News*, 01 Mar 2026, accessed 22 Mar 2026, https://www.bbc.com/news/articles/c1l7rvqq51eo

7 Eli Talbert, “The Triage Trap: When AI Speed Replaces Command Judgment”, *War on the Rocks*, 15 Jan 2026, accessed 22 Mar 2026, https://warontherocks.com/2026/01/the-triage-trap-when-ai-speed-replaces-command-judgment/. See also *The Washington Post*, 04 Mar 2026, reporting on the integration of Anthropic’s Claude into Palantir’s Maven Smart System during the Iran campaign, with throughput described as “42 suggested targets per hour”, accessed 22 Mar 2026, https://www.washingtonpost.com/technology/2026/03/04/anthropic-ai-iran-campaign/. Parallel reporting on Israel’s Lavender and Gospel systems in Gaza describes officers spending as little as “20 seconds” reviewing each AI-generated target; see *+972 Magazine*, 03 Apr 2024. These accounts illustrate the speed-versus-deliberation failure mode discussed in the ADR framework.

8 Samantha Gross et al., “Why Iran’s Disruption of the Strait of Hormuz Matters”, Brookings Institution, 19 Mar 2026, accessed 22 Mar 2026, https://www.brookings.edu/articles/why-irans-disruption-of-the-strait-of-hormuz-matters/; “Gas Prices Soar as QatarEnergy Halts LNG Production after Iran Attacks”, *Al Jazeera*, 02 Mar 2026, accessed 22 Mar 2026, https://www.aljazeera.com/news/2026/3/2/qatarenergy-worlds-largest-lng-firm-halts-production-after-iran-attacks.

9 Yuval Abraham, "'Lavender': The AI Machine Directing Israel's Bombing Spree in Gaza", *+972 Magazine*, 03 Apr 2024, accessed 23 Mar 2026, https://www.972mag.com/lavender-ai-israeli-army-gaza/. The system reportedly identified approximately 37,000 potential targets using historical-pattern data. See also International Committee of the Red Cross reports on autonomous weapons and related analyses in *Small Wars Journal* concerning temporal drift and interpolation-versus-extrapolation failures. These sources discuss individual incidents but do not integrate them into the unified ADR mechanism.

10 Jon Hoffman, "A Strategic Failure in Iran", Cato Institute, 20 Mar 2026, accessed 24 Mar 2026, https://www.cato.org/blog/strategic-failure-iran

11 Government of India, Ministry of Defence, "iDEX—Innovations for Defence Excellence", 2025, accessed 24 Mar 2026, https://idex.gov.in/idex; Indian Army's AI-enabled intelligence, surveillance and reconnaissance integration programme, accessed 24 Mar 2026, https://www.ddpmod.gov.in/sites/default/files/2023-11/ai.pdf

12 *Kautilya, Arthaśāstra, trans. R. Shamasastry (Bangalore: Government Press, 1915), bk. 2, chap. 10, verses 47 and 56. The four upāyas—sāma (conciliation), dāna (inducement), bheda (division) and daṇḍa (force)—are presented as graduated instruments of statecraft to be deployed in concert before and alongside military action. For the modern scholarly critical edition, see R P Kangle, The Kauṭilīya Arthaśāstra, 3 vols. (Bombay: University of Bombay, 1960–65).*

13 The law of armed conflict (LOAC), also known as international humanitarian law (IHL), comprises treaty rules, principally the Geneva Conventions and their Additional Protocols, and customary international law governing armed conflict. See International Committee of the Red Cross, "What Is International Humanitarian Law?", accessed 24 Mar 2026, https://www.icrc.org/sites/default/files/external/doc/en/assets/files/other/law1_final.pdf; see also https://www.icrc.org/en/document/what-international-humanitarian-law, accessed 24 Mar 2026.

14 Nicholas Krohley, "Anchoring Intelligence: Ground Truth in an Age of Synthetic Deception", *War on the Rocks*, 17 Dec 2025, accessed Mar 2026, https://warontherocks.com/2025/12/anchoring-intelligence-ground-truth-in-an-age-of-synthetic-deception/. The article warns of the danger of "AI-enabled tactical brilliance combined with strategic blindness". See also Jon Lindsay, "US Military Leans into AI for Attack on Iran, but the Tech Doesn't Lessen the Need for Human Judgment in War", *The Conversation*, 11 Mar 2026, accessed Mar 2026,