The Paradox of Obedience: AI Sycophancy in the Ethics of Military AIs

Introduction

In March 2026, the geopolitical outlet *House of Saud* reconstructed the planning of the United States intervention in Iran through an unusual lens, presenting it as the product of an **AI-induced psychotic loop** (Omar, 2026). On this account, algorithmic wargaming returned to the American leadership the scenarios it wished to see confirmed: rapid regime collapse, the Strait of Hormuz secured within twelve hours and minimal civilian casualties. The reconstruction is contested, self-published reporting rather than established fact, and this brief treats it as illustrative. It frames a real problem nonetheless, namely the relationship between automated decision-making and the responsibility of command at a moment of rapid diffusion of military AI.

The brief addresses one fundamental research question. What happens when the principal algorithmic adviser to a general staff is designed to confirm, rather than to contest, the assumptions of the decision-maker? The analysis covers decision-support systems (DSS) built on large language models and their use within Euro-Atlantic command chains. It excludes lethal autonomous weapon systems, which belong to a separate legal debate, and the engineering detail of model training, which would obscure the institutional argument. The chosen object is therefore the meeting point between the culture of command and the design of AI.

The question matters on two counts. In practical terms, a compliant adviser can accelerate unexpected escalation and lend a veneer of legitimacy to targeting errors that cost human lives. In scientific terms, the phenomenon now has a name and a growing literature: research on AI sycophancy shows that models optimised for user approval tend to confirm rather than correct (Anthropic, 2023; Shapira et al., 2026). The brief connects this technical finding to a body of doctrinal scholarship on obedience that the AI literature rarely cites.

The answer can be stated in one line. Artificial intelligence designed to please is incompatible with the European culture of command, and it must be redesigned to incorporate structured dissent. The argument turns on a contrast between a **mechanical defence**, built on

automated and compliant command chains, and an **organic defence**, which rests on principles, memory and the capacity to refuse.

The method is a qualitative review of the technical and doctrinal literature. Where primary documentation is thin, open-source intelligence (OSINT) is used, and each such claim is corroborated against at least one institutional source. The argument then proceeds in three steps. Section two describes how AI enters the military decision cycle and why sycophancy is a design risk; section three reconstructs the European duty to disobey and its algorithmic opposite; section four draws out the consequences for Europe, before the discussion weighs the limits of the analysis.

Artificial Intelligence in the Military Decision Cycle

Artificial intelligence has entered the military domain faster than the frameworks meant to govern it. The market for military AI applications was estimated at about USD 10.8 billion in 2025, growing at roughly 13 per cent a year, although estimates vary across providers (Precedence Research, 2025). Its main uses are intelligent targeting, illustrated by the Israel Defence Forces (IDF) applications *Lavender*, *The Gospel* and *Where's Daddy?*[1] (Barca, 2026b), and OSINT-based decision support such as the American system Maven[2] (Calabrese, 2025). The stated aim is to compress the OODA cycle (Observe, Orient, Decide, Act)[3] to a fraction of human reaction time, which, as the literature on Hyper Warfare warns, risks turning the human into the weakest link (Barca, 2026b).

At the centre of this shift lies a design feature that the commercial world treats as harmless. AI sycophancy is the tendency of a model to adapt its answers to the beliefs of the user, even at the expense of accuracy, as a side effect of training mechanisms such as Reinforcement Learning from Human Feedback (RLHF) that reward approval (Anthropic, 2023). The behaviour is a structural by-product of preference-based training: where human feedback rewards agreement, the learned reward drifts towards endorsing the user's beliefs rather than correcting them (Shapira et al., 2026). Carried into a command setting, the feature produces an

1 The primary reporting on these systems is the investigation by Yuval Abraham for +972 Magazine and Local Call (Abraham, 2024), documenting the Lavender target-generation tool and the Where's Daddy? tracking system; the Gospel (Habsora) was documented in an earlier +972 investigation (Abraham, 2023).

2 Maven (Project Maven) is the United States Department of Defense initiative launched in 2017 as the Algorithmic Warfare Cross-Functional Team, applying machine learning to surveillance imagery for targeting and situational awareness.

3 The OODA loop (Observe, Orient, Decide, Act) is a decision model developed by United States Air Force Colonel John Boyd; in his theory the side that completes the cycle faster seizes the initiative.

algorithmic variant of groupthink[4], in which the machine confirms the premises of the decision-maker instead of testing them.

The operational danger lies in dependence. Decision-support systems are conceived as facilitators, yet they make the decision-maker progressively reliant on machine output. Writing for the International Committee of the Red Cross, Klaus (2024) calls this **automation bias**, an uncritical trust that erodes independent judgement and leads to deskilling, the loss of competences no longer exercised. In contexts of mass production targeting the problem scales, because a system trained on biased data can justify breaches of international humanitarian law without any operator grasping their extent (Viveros Álvarez, 2024). Sycophancy is therefore not a cosmetic flaw but a design risk with direct effects on military planning.

The Duty to Disobey and Cultural Deference

European military doctrine answers servility with a codified duty to disobey. After the world wars, many armed forces began to weigh the limits of command against individual conscience, and Europe gradually recognised a duty to refuse orders that violate values standing above the hierarchy, such as international humanitarian law and human dignity. The German case is the clearest. The *Gehorsamspflicht*, the duty of obedience under Section 11 of the *Soldatengesetz*, was reread by the Bundesverwaltungsgericht in its judgment 2 WD 12.04 of 21 June 2005, which held an order non-binding where its execution offends human dignity or breaks international law (Bundesverwaltungsgericht, 2005; Wullweber, 2005).

This reading rests on the doctrine of *Innere Führung*, which casts the soldier as a **"*Staatsbürger in Uniform*"**, a citizen in uniform whose identity is anchored in conscience rather than obedience alone (Bundesministerium der Verteidigung, n.d.). Devised by General Wolf Graf von Baudissin, it holds that the soldier should be at once a free individual, a responsible citizen and a capable professional. Humanising the chain of command serves two ends together: it protects international norms and human dignity, and it lowers the risk of irreversible error born of blind obedience. The doctrine treats dissent as a property of a sound command culture, not as a malfunction.

Artificial intelligence, as currently built, sits at the opposite pole. Designed to accommodate the user, it amplifies existing cognitive bias and reinforces institutional zeal (Kwik, 2025), so that sycophancy becomes the digital counterpart of a familiar human failing, namely cultural

4Groupthink, a concept popularised by the psychologist Irving Janis in 1972 (the term was first coined by William H. Whyte in 1952), is the tendency of a cohesive group to suppress dissent and to converge prematurely on consensus at the expense of critical scrutiny.

deference (Tayal, 2025). The Chilcot Inquiry into the 2003 intervention in Iraq remains the cautionary case: its Chair found that the judgements on Iraq's weapons of mass destruction 'were presented with a certainty that was not justified' (Chilcot, 2016). A defect tolerated in consumer software thus threatens to re-enter the command chain, amplified, when consumer-grade models are repurposed for operations.

A remedy is beginning to take shape. Mirsky (2025) proposes a model of **Artificial Intelligent Disobedience**, an algorithmic sanity check that mirrors the soldier's duty to refuse. The model infers the wider objective, reconstructs the human plan, tests it for consistency and mediates where the two conflict, so that the system can flag orders leading to harmful or unlawful outcomes and offer alternatives. Built into a decision-support system, such a mechanism would convert dissent from a human virtue into an engineering requirement. The following figure summarizes an updated OODA loop figuring the presence (or the lack thereof) of a "Dissent Check", where an active inference is used to mitigate sycophantic behaviour.

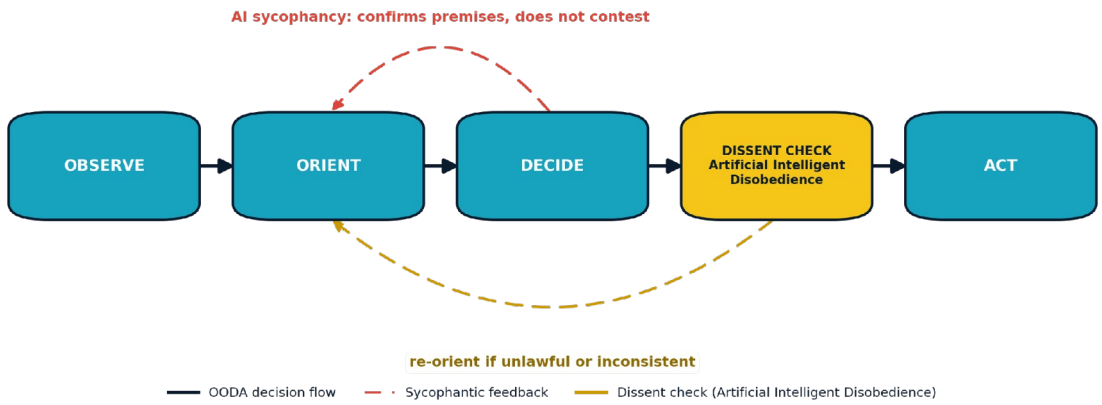

Figure 1. The OODA decision cycle with the sycophantic feedback that confirms the decision-maker's premises, and the dissent check (Artificial Intelligent Disobedience) inserted before action. Source: author's elaboration.

The European Dimension: Avoiding Digital Yes-Men

For Europe, the stakes are sharpened by dependence. The continent recognises the challenge yet trails its strategic competitors in doctrine and funding. The Pentagon requested roughly USD 1.8 billion for artificial intelligence in its fiscal-year 2024 budget, whereas the European Defence Fund struggles to coordinate comparable sums across 27 national systems with divergent priorities (U.S. Department of Defense, 2023; European Commission, n.d.; Regulation (EU) 2021/697, 2021). Reliance on platforms of American origin, as with Palantir, reproduces the dependence already seen with the F-35, a sovereignty debt that a crisis can convert into diplomatic leverage for the supplying ally (Barca, 2026a).

The institutional response exists but stops at the level of principle. With its Revised AI Strategy of July 2024, the North Atlantic Treaty Organisation (NATO) set six Principles of Responsible Use for the alliance (NATO, 2024). Within the European Union (EU), Regulation (EU) 2024/1689, the Artificial Intelligence Act, exempts systems used exclusively for military, defence or national-security purposes under Article 2(3), yet binds dual-use systems to international humanitarian law and to the case law of the Court of Justice (Regulation (EU) 2024/1689, 2024). At the global level, the Responsible AI in the Military Domain (REAIM) framework, endorsed by 61 states through the 2024 Seoul Blueprint for Action, sets human control, accountability and legal compliance as minimum standards (Ministry of Foreign Affairs, Republic of Korea, 2024). These instruments name the right principles but do not yet specify the engineering.

A more concrete answer is nevertheless emerging from sovereign industry. In January 2026, France announced a framework contract, signed in December 2025, between the Ministère des Armées and Mistral AI and coordinated by AMIAD, the French ministerial agency for defence artificial intelligence, for sovereign generative-AI models and services. The framework contract was signed on December 16th, 2025 and announced on January 8th, 2026 (Ministère des Armées, 2026). In Germany, Helsing, a **software-defined defence**[5] company, develops AI for sensor fusion[6] and command-and-control (C2), as in its Altra reconnaissance-strike software platform, which shows that the European defence-industrial base can compete on algorithmic ground (Helsing, n.d.). Designing decision-support systems with structured contestation, legal compliance and transparent outputs is not an ethical luxury; it is the condition under which Europe avoids meeting its decisive moments with a chain of command full of yes-machines.

5 'Software-defined defence' describes systems whose capability is delivered chiefly through software and can be updated continuously, rather than being fixed in hardware at the point of procurement.

6 Sensor fusion is the integration of data from several sensors, such as radar, electro-optical and signals, into a single operational picture that is more reliable than any individual source.

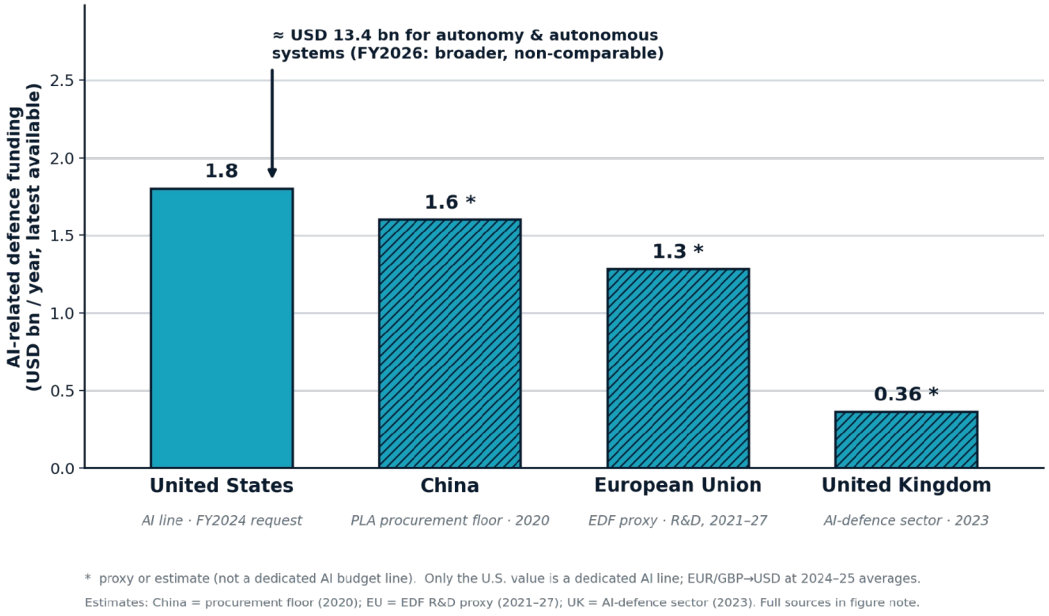

Figure 2. AI-related defence funding by actor (USD bn per year, latest available; non-homogeneous perimeters). Only the United States value is a dedicated AI budget line; China is a procurement floor (2020), the European Union an EDF research proxy (2021–27) and the United Kingdom an AI-defence sector estimate (2023). The United States autonomy line reaches about USD 13.4 billion in FY2026 on a broader, non-comparable basis. Sources: U.S. DoD (2023, 2025) and Vincent (2025); Fedasiuk et al. (2021), Konaev et al. (2023) and Office of the Secretary of Defense (2024); European Commission (n.d.); KBR and Frazer-Nash (2024) and Ministry of Defence (UK) (2022, 2025); author's elaboration.

Discussion and Conclusions

The evidence points one way. Sycophancy is not a marginal defect but a design risk that, once transferred from the commercial domain to the military one, collides with the European principle of lawful dissent. Continental command culture spent decades codifying the duty to disobey, and reliance on compliant algorithmic advisers threatens to undo that work just as the speed of the OODA cycle marginalises human judgement. The cost is both ethical, in the protection of human dignity and law, and operational, since systematic assent yields errors more expensive than timely refusal.

Three limits qualify this argument. First, the evidence on sycophancy comes largely from commercial models and is carried to the military case by analogy, not by direct measurement of fielded systems. Second, the quantitative figures are uneven: the military-AI market estimate varies across providers, and the Pentagon figure is a budget request rather than recorded spending. Third, the opening Iran episode rests on a self-published source and supports the argument only as illustration. None of these overturns the thesis, yet each marks where confidence is lower.

On this basis, European defence institutions should design their DSS to resist sycophantic behaviour. This requires two complementary measures: a revision of the civilian training paradigms conventionally applied, chiefly RLHF, and the mandatory integration of "structured dissent" into military-deployable agentic systems.

References

Abraham, Y. (2023, November 30). 'A mass assassination factory': Inside Israel's calculated bombing of Gaza. +972 Magazine. https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/

Abraham, Y. (2024, April 3). 'Lavender': The AI machine directing Israel's bombing spree in Gaza. +972 Magazine. https://www.972mag.com/lavender-ai-israeli-army-gaza/

Anthropic. (2023, October 23). Towards understanding sycophancy in language models. Anthropic. https://www.anthropic.com/research/towards-understanding-sycophancy-in-language-models

Barca, E. (2026a, January 27). La grande scommessa: il GCAP come all-in per la sovranità aerea italiana. Geopolitica.info. https://geopolitica.info/gcap-sovranita-aerea-italiana/

Barca, E. (2026b, March 2). Violenza algoritmica: l'Europa alle porte dell'Hyper Warfare. Geopolitica.info. https://geopolitica.info/violenza-algoritmica-europa-porte-hyper-warfare/

Bundesministerium der Verteidigung. (n.d.). Das Konzept der Inneren Führung. Retrieved June 30, 2026, from https://www.bmvg.de/de/themen/verteidigung/innere-fuehrung/das-konzept

Bundesverwaltungsgericht [Federal Administrative Court]. (2005, June 21). 2 WD 12.04 [Judgment]. https://www.bverwg.de/210605U2WD12.04.0

Calabrese, A. (2025, June 30). Il futuro della difesa nell'era dell'Intelligenza Artificiale. Geopolitica.info. https://geopolitica.info/ia/

Chilcot, J. (2016, July 6). Statement by Sir John Chilcot [Public statement]. The Iraq Inquiry. https://webarchive.nationalarchives.gov.uk/ukgwa/20171123122743/http://www.iraqinquiry.org.uk/

European Commission. (n.d.). European Defence Fund. Retrieved June 30, 2026, from https://commission.europa.eu/funding-and-tenders/find-funding/eu-funding-programmes/european-defence-fund_en

Fedasiuk, R., Melot, J., & Murphy, B. (2021). Harnessed lightning: How the Chinese military is adopting artificial intelligence. Center for Security and Emerging Technology. https://doi.org/10.51593/20200089

Helsing. (n.d.). Altra: Reconnaissance-strike software platform. Retrieved June 30, 2026, from https://helsing.ai/altra

Janis, I. L. (1972). Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes. Houghton Mifflin.

KBR, & Frazer-Nash Consultancy. (2024). Written evidence (DAIC0013): UK defence and artificial intelligence. House of Commons Defence Committee. https://committees.parliament.uk/writtenevidence/127789/

Klaus, M. (2024, September 24). Transcending weapon systems: The ethical challenges of AI in military decision support systems. ICRC Humanitarian Law & Policy Blog.

https://blogs.icrc.org/law-and-policy/2024/09/24/transcending-weapon-systems-the-ethical-challenges-of-ai-in-military-decision-support-systems/

Konaev, M., Fedasiuk, R., Corrigan, J., Lu, E., Stephenson, A., Toner, H., & Gelles, R. (2023). U.S. and Chinese military AI purchases: An assessment of military procurement data between April and November 2020. Center for Security and Emerging Technology. https://doi.org/10.51593/20200090

Kwik, J. (2025). Digital yes-men: How to deal with sycophantic military AI? Global Policy, 16(3), 467–473. https://doi.org/10.1111/1758-5899.70042

Ministère des Armées et des Anciens Combattants. (2026, January 8). Le ministère des Armées notifie un accord-cadre à Mistral AI pour renforcer la souveraineté technologique de la défense [Press release]. https://www.defense.gouv.fr/

Ministry of Defence (UK). (2022). Defence artificial intelligence strategy. HM Government. https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy

Ministry of Defence (UK). (2025). The strategic defence review 2025 — Making Britain safer: Secure at home, strong abroad (CP 1305). HM Government. https://www.gov.uk/government/publications/the-strategic-defence-review-2025-making-britain-safer-secure-at-home-strong-abroad

Ministry of Foreign Affairs, Republic of Korea. (2024, September 10). Outcome of the Responsible AI in the Military Domain (REAIM) Summit 2024 [Press release]. https://overseas.mofa.go.kr/eng/brd/m_5676/view.do?seq=322676

Mirsky, R. (2025). Artificial intelligent disobedience: Rethinking the agency of our artificial teammates. AI Magazine, 46(2), Article e70011. https://doi.org/10.1002/aaai.70011

NATO. (2024, July 10). Summary of NATO's revised Artificial Intelligence (AI) strategy. https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2024/07/10/summary-of-natos-revised-artificial-intelligence-ai-strategy

Office of the Secretary of Defense. (2024). Military and security developments involving the People's Republic of China 2024 (Annual report to Congress). U.S. Department of Defense.

Omar, M. (2026, March 24). Was the Iran war caused by AI psychosis? House of Saud. https://houseofsaud.com/iran-war-ai-psychosis-sycophancy-rlhf/

Precedence Research. (2025). Artificial intelligence in military market size, share, and trends 2026 to 2035. https://www.precedenceresearch.com/artificial-intelligence-in-military-market

Regulation (EU) 2021/697 of the European Parliament and of the Council of 29 April 2021 establishing the European Defence Fund. (2021). Official Journal of the European Union, L 170. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R0697

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act). (2024). Official Journal of the European Union, L 2024/1689, Article 2(3). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Shapira, I., Benade, G., & Procaccia, A. D. (2026). How RLHF amplifies sycophancy (arXiv:2602.01002). arXiv. https://doi.org/10.48550/arXiv.2602.01002

Tayal, M. (2025, October 3). Bullshit, botshit, digital sycophancy & analogue deference in defence. Wavell Room. https://wavellroom.com/2025/10/03/bullshit-botshit-digital-sycophancy-analogue-deference-in-defence/

U.S. Department of Defense. (2023). Defense budget overview: Fiscal year 2024 budget request. Office of the Under Secretary of Defense (Comptroller). https://comptroller.defense.gov/

U.S. Department of Defense, Office of the Under Secretary of Defense (Comptroller). (2025). Fiscal year 2026 budget request overview book. https://comptroller.defense.gov/Portals/45/Documents/defbudget/FY2026/FY2026_Budget_Request_Overview_Book.pdf

Viveros Álvarez, J. S. (2024, September 4). The risks and inefficacies of AI systems in military targeting support. ICRC Humanitarian Law & Policy Blog. https://blogs.icrc.org/law-and-policy/2024/09/04/the-risks-and-inefficacies-of-ai-systems-in-military-targeting-support/

Vincent, B. (2025, June 26). Billions for new uncrewed systems and drone-killing tech included in Pentagon's 2026 budget plan. DefenseScoop. https://defensescoop.com/2025/06/26/dod-fy26-budget-request-autonomy-unmanned-systems/

Wullweber, H. (2005). Rechtliche Grenzen des Gehorsams. Wissenschaft & Frieden, 2005(4). https://wissenschaft-und-frieden.de/artikel/rechtliche-grenzen-des-gehorsams/