Buchbesprechungen

***Deeks, Ashley S.*: The Double Black Box. National Security, Artificial Intelligence, and the Struggle for Democratic Accountability**. Oxford UK: Oxford University Press, 2025. ISBN 978-0-19-752090-1. x, 260 pp. £18.99

In February 2026, a dispute between the Department of War (DoW) and the US-based Artificial Intelligence (AI) company Anthropic, once again underscored how little is publicly known about the use of AI in warfare. The disagreement centred on Anthropic's insistence that Claude, its large language model, cannot be used safely or reliably in fully autonomous weapons or for mass domestic surveillance.[1] The DoW refused to accommodate Anthropic's demands, designating the company a supply chain risk instead.[2] Beyond a few heated exchanges, however, the public record quickly becomes thin: it remains unclear how the United States deploys Claude; how it interfaces with the broader targeting support system used by the DoW, Palantir's Maven System; what safeguards, if any, are operative; and what it would mean in practice for such use to satisfy legal requirements. Disconcertingly, shortly after the dispute, reports suggested that Claude may have been used in US strikes on Iran.[3] This has prompted speculation as to whether this frontier model played a role in the specific attack on the Shajareh Tayyebeh Elementary School in Minab, which killed at least 175 people.[4] At the time of writing, key facts remain elusive.

Although these events occurred after the publication of Ashley Deeks' 'The Double Black Box: National Security, Artificial Intelligence, and the Struggle for Democratic Accountability', they proved difficult to set aside while reading it. The book is written at a time when the need to understand and penetrate the layers of secrecy surrounding AI-enabled military operations feels particularly acute. Deeks elegantly guides the reader through what she terms the 'double black box': the familiar opaqueness of national security

1 'Statement from Dario Amodei on our discussions with the Department of War,' <https://www.anthropic.com/news/statement-department-of-war>, last access 9 June 2026.

2 Lily Jamali and Kali Hays, 'Anthropic Vows to Sue Pentagon Over Supply Chain Risk Label', BBC News, 6 March 2026, <https://www.bbc.com/news/articles/cn5g3z3xe65o>, last access 22 May 2026.

3 'Anthropic's AI tool Claude Central to US Campaign in Iran, Amid a Bitter Feud', The Washington Post, 4 March 2026, <https://www.washingtonpost.com/technology/2026/03/04/anthropic-ai-iran-campaign/>, last access 22 May 2026; Ed Pilkington, 'US Military Reportedly Used Claude in Iran Strikes Despite Trump's Ban', The Guardian, 1 March 2026, <https://www.theguardian.com/technology/2026/mar/01/claude-anthropic-iran-strikes-us-military>, last access 22 May 2026.

4 See, for example, Akshaya Kumar, 'Was the Attack on an Iranian Primary School a War Crime?', Lawfare, 16 April 2026.

DOI 10.17104/0044-2348-2026-3-999

decision-making, compounded by the additional inscrutability introduced by AI systems. The author's central claim is straightforward: as these layers of opacity converge, 'a range of actors who serve as surrogates for the US public – who already seek, and sometimes fail, to check the Executive in the national security space – will have an increasingly hard time doing so' (p. 11).[5] Deeks' diagnosis is as persuasive as it is disquieting.

'The Double Black Box' considers the national security applications of AI, including target nomination, the automation of kinetic and cyber operations, border-related arrest decisions, and counterterrorism activities. Paradoxically, although these uses present some of the most significant risks, they remain among the least regulated. While jurisdictions such as the United States and the European Union have developed domestic frameworks governing civilian AI uses, military applications continue to operate largely outside dedicated regulatory regimes, whether at the international, EU, or national level. As Sullivan observes, 'even modest international efforts to legally constrain military AI appear dead on arrival, while domestic regulatory frameworks show little appetite for constraining national military initiatives'.[6]

Much of the scholarly debate on military AI and opacity has focused on its implications for users' understanding of AI-generated outputs, for example, whether individuals can comprehend or subsequently explain the basis of a system's recommendation.[7] Relatedly, the concern with AI opacity has also led to an important strand of literature on meaningful human control.[8] Deeks' main contribution is the shift from the challenges faced by end-users individually to societal questions of democratic accountability more broadly. She convincingly demonstrates that opacity in national security contexts is more complex, layered, and compounded than commonly assumed. Its implications are correspondingly more profound, with the potential to systematically erode what she calls 'public law values': (i) legal compliance,

5 Secrecy surrogates in this context refers to 'actors who are given access to secret information that average US citizens are not and who can improve secret executive operations and help mitigate abuses' (p. 62).

6 Scott Sullivan, 'Military AI as "Abnormal" Technology,' Lawfare, 12 March 2026.

7 See, for example, Arthur Holland Michel, 'The Black Box, Unlocked: Predictability and Understandability in Military AI', Geneva, Switzerland: United Nations Institute for Disarmament Research (2020), doi: 10.37559/SecTec/20/AI1, 9-11; Nathan G. Wood, 'Explainable AI in the Military Domain', Ethics and Information Technology 26 (2024), 29 <https://doi.org/10.1007/s10676-024-09762-w>.

8 See, for example, Michel (n. 7), p. 13; Bérénice Boutin and Taylor K. Woodcock, 'Aspects of Realizing (Meaningful) Human Control: a Legal Perspective' in: Robin Geiß and Henning Lahmann, Research Handbook on Warfare and Artificial Intelligence, (Edward Elgar Publishing 2024), <https://doi.org/10.4337/9781800377400.00016>.

(ii) competence and rational decision-making, (iii) accountability for decisions made, and (iv) justification or reason-giving to appropriate actors (p. 13).

The book will be of interest to scholars, policy-makers and the general public alike – to all those concerned with the impact of AI on the broader socio-political processes within domestic systems and in international relations. As Deeks highlights, what is at stake in the double black box is the potential for the concealment of mistakes and abuses, the avoidance of accountability, and the perpetuation of problematic policies or systems without any critique (pp. 31-38). In itself, this is not a new problem, as governments have obscured troubling practices for decades. However, AI's integration can make it harder to detect and correct errors before they cause serious damage (p. 107). This is largely because AI's opacity can prevent those involved from recognising unfair, inaccurate, or misleading outputs relied upon in the national security context, for instance, for targeting or detention. Indeed, the nature of AI systems impedes not only public oversight but also the ability of Congress, the judiciary, journalists, and even officials within the Executive to pose the necessary questions or conduct scrutiny that might prevent costly, widespread mistakes.

A noteworthy element of Deeks' argument is that opacity in AI-enabled processes is not only a feature of the model's complexity, as it is commonly perceived. It is also the speed of AI-enabled military operations, especially in relation to the cyber domain, that further exacerbates the black box problem. This is because the tempo at which AI-enabled decisions and effects occur outpaces the capacity of users to report concerns to oversight bodies, in turn also undermining the ability of competent authorities to intervene to prevent violations in a timely manner (pp. 128, 146-147). The argument could further be strengthened by recognising that, in addition to speed, the potential for large volume of AI outputs also contributes to the problem of opacity, as users and institutions become practically unable to meaningfully review or verify such vast quantities of AI recommendations.[9] Taken together, AI contributes to the opaqueness of operations not only because of the inscrutability of its models, but because of the speed and volume of outputs. As a result, as Deeks argues, the conditions created by AI dramatically narrow the space for effective oversight, prevention of errors, and democratic accountability.

In my view, it is important to acknowledge, as Deeks does in passing, that AI companies and the choices they make in the design and development phases of these technologies significantly exacerbate the double black box. For example, designers intentionally develop complex AI models or decline to reveal

9 Jessica Dorsey and Marta Bo, 'Symposium on Military AI and the Law of Armed Conflict: The "Need" for Speed – The Cost of Unregulated AI Decision-Support Systems to Civilians', OpinioJuris, 4 April 2024.

DOI 10.17104/0044-2348-2026-3-999

the datasets on which the systems were trained (pp. 61, 81, 193). They may also be non-transparent about the assumptions underpinning the model. As other authors have also noticed, the opacity of AI systems is a choice that advances commercial interests but undermines the ability of government agencies to meaningfully understand and review the systems they acquire.[10]

The strategic role of secrecy and how actors exploit it to exercise power resonate with agnotology studies in international relations. In this regard, the work of Aradau and Huysmans is of particular relevance, in which they examine how knowledge surrounding data and algorithmic practices is produced.[11] They argue that the opacity of those practices allows actors to assemble authority and power, by rendering only certain forms of knowledge credible.[12] Those denied access to such knowledge – for example, technical information about AI systems – are effectively barred from advancing meaningful critique of those practices. These dynamics of power exemplify the 'double black box': government actors and increasingly private technology companies, exploit the opacity of AI-enabled operations and institutionalise that opacity to consolidate and advance their power.

A common response to the black box problem is to seek transparency. However, 'opening up' the black box is not, in itself, a sufficient solution to concerns surrounding opacity. Or, as Deeks observes, it may in practice prove both insufficient and uninformative (p. 74). This is especially the case with respect to the algorithmic 'black box'; even where source code is made accessible, the scale and complexity of contemporary models – often comprising millions of lines of code – render it exceedingly difficult to reconstruct how the model functions, to identify its underlying assumptions, or to trace the origins of specific errors.[13] These challenges are further compounded in contexts involving interoperable systems or when the code is AI-generated and thus introduces additional layers of opacity.[14]

10 Roy Lindelauf and Herwin Meerveld, 'Building Trust in Military AI Starts with Opening the Black Box', War on the Rocks, 12 August 2025. Rudin also observed a similar dynamic by noting that, in some circumstances, similar performance could be achieved by employing interpretable algorithmic models, but that companies are more interested in opaque models as those can be more easily made proprietary. Cynthia Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead', Nature Machine Intelligence (2019), 206-215, <https://doi.org/10.1038/s42256-019-0048-x>.

11 Claudia Aradau and Jef Huysmans, 'Assembling Credibility: Knowledge, Method and Critique in Times of "Post-Truth"', Security Dialogue 50 (2019), 40-58, <https://doi.org/10.1177/0967010618788996>.

12 Aradau and Huysmans (n. 11).

13 Michel (n. 7), 9.

14 See, for example, Daniel Trusilo, 'Autonomous AI Systems in Conflict: Emergent Behavior and Its Impact on Predictability and Reliability', Journal of Military Ethics 22 (2023), 2-17, <https://doi.org/10.1080/15027570.2023.2213985>.

Fortunately, the book offers elegant solutions that do not necessarily involve 'opening up' the proverbial 'black box'. Deeks instead argues that actors within the national security apparatus should, at a minimum, be able to explain why particular AI systems were acquired or developed, and what measures were adopted to ensure their legally compliant use. Reliance on generic or abstract justifications – such as a broad appeal to the utility of algorithmic recommendations – is unlikely to satisfy public scrutiny or accountability demands (p. 118). Instead, 'strategic transparency', the effort to convey to the public the difficult trade-offs related to the use of AI capabilities, explaining 'how the tools are used, by whom, and pursuant to what legal rules' (p. 170), offer the advantages of engaging with the critics, diffusing objections, lowering public anxiety or even obtaining public buy-in, and in doing so, hopefully, learning something from the process that can improve the AI-assisted decision-making processes (pp. 170-175). Deeks' suggestions are of great value including to the current US administration. The Anthropic-DoW dispute, for example, was a missed opportunity for the government to explain how it operationalises legal obligations in its use of AI, how it verifies and assures the reliability of large language models, or the measures it takes to prevent abuses, especially in the ongoing armed conflicts in which the US is involved.

Another proposal described in the book is the introduction of a legislative requirement for explainable AI (xAI), which could have the effect of disincentivising the deployment of highly complex models, including large language models, where their opacity renders them insufficiently explainable (p. 150). This entails, first, that companies supplying such systems are able to provide meaningful explanations of their functioning, and second, that government actors retain the capacity to articulate and transmit those explanations to relevant audiences, including end-users and, subsequently, those tasked with assessing or contesting the legality of their use. The calls for explainable AI, however, require refinement, for example, in specifying the type of explanations required, the audiences to whom they are addressed, and the appropriate structure and format such explanations should take (p. 154).

I further agree with Deeks' argument that the double black box challenge cannot be solved without greater AI literacy amongst those organs and institutions that are tasked with oversight or monitoring of the AI-enabled practices. What is meant by AI literacy is not necessarily an ability to inspect the models themselves, but rather the capacity to ask appropriate questions and seek justifications for the use of AI (p. 127). Similarly, officers working at other national security agencies or even foreign security partners that have access to information about classified programs would benefit from the necessary training or prior expertise to be able to serve in their role of non-

DOI 10.17104/0044-2348-2026-3-999

traditional checks (p. 158). This includes understanding the risks involved in the use of AI, common concerns around biases, errors, performance fluctuations, opacity, and vulnerability to adversarial attacks. In my experience, the US and other governments are increasingly investing resources in training that will strengthen these skills.

There is, however, one proposal in the book that I remain unconvinced by. While Deeks acknowledges that AI companies have a commercial interest in creating opaque models, she also places great hope in their role as secrecy surrogates, that is, actors that are able to identify, report, and mitigate abuses of national security actors (p. 62). She argues that, through the provision of technologies and related services, AI companies become privy to intimate knowledge about national security practices, putting them in a unique position to act.

In some ways, Anthropic's demands to exclude Claude from use in autonomous weapons or mass surveillance can be seen precisely as a successful example of how AI companies can play the role of secrecy surrogates. Indeed, the Guardian even reported that Anthropic 'is functioning as one of the only checks on what appears to be the military's expansive desires for weaponizing AI'.[15] At the same time, Anthropic was not able to uphold the red lines it proposed. As Deeks predicted, any refusal to cooperate with the government by one company results in its replacement (p. 195). In this case, several days after the dispute was made public, OpenAI has already signed an agreement with the DoW to offer its large language model as an alternative.[16] Even more importantly, in my view, it must not be underestimated how entangled AI companies are in the contemporary national security practices. The commercially-driven choices in AI design and development generate structural dependencies of the national security actors on the private sector and reliance on their expertise.[17] As also shown through the Anthropic-DoW dispute, even when disagreements emerge, it takes time for the national security infrastructure to even disconnect or rather untangle its ties with the company. Rather than contributing to unveiling violations, AI companies are deeply entangled and complicit in the military and law enforcement operations.

15 Blake Montgomery, 'Datacenters Are Becoming a Target in Warfare for the First Time', The Guardian, 10 March 2026.

16 'Our agreement with the Department of War', OpenAI, <https://openai.com/index/our-agreement-with-the-department-of-war/>, last access 22 May 2026.

17 Marijn Hoijtink and Jasper van der Kist, 'Platforms on the Frontline: The Rise of the Platform Model in Defense Tech', OpinioJuris, 11 February 2025; Marijn Hoijtink, Jasper van der Kist and Laszlo Steinwärder, 'The Rise of the Tech Oligarchy and Its Military Entanglements: a Platform Approach', Sciences as Culture (2026), 1-15 (7-8), <https://doi.org/10.1080/09505431.2026.2666079>.

 DOI 10.17104/0044-2348-2026-3-999

'The Double Black Box' is centered on the US constitutional framework, including its mechanisms of unveiling secretive national security practices through the Executive (pp. 50-56), Congress (pp. 37-44), judicial bodies (pp. 44-49), and whistleblowers (pp. 56-59). However, the lessons learned from this book are going to resonate with an international audience for at least two reasons. Firstly, all democratic States committed to the rule of law should strive to ensure that the use of AI systems does not deepen the veil of secrecy surrounding them. As mentioned by Deeks, we should be cautious 'about the ways in which AI's characteristics affect the ability of actors outside (and inside!) democratic States, including the United States, to oversee, check, and challenge executive national security activities' (p. 18). Given the many dilemmas that accompany the use of AI systems in national security contexts, mechanisms must be in place to check the developments and uses of technologies with such high risks.

Secondly, the US is an actor with a particularly high involvement in armed conflicts around the world and significant AI capabilities. For those who may fall victim to the use of AI systems in US military operations or for other States who wish to ensure respect for international humanitarian law, it is also important that there is greater clarity with respect to how the United States is employing AI capabilities and what checks and balances it has in place to prevent unlawful uses.

Klaudia Klonowska, Sciences Po Paris, Paris, France

DOI 10.17104/0044-2348-2026-3-999