Upstream Dual-Use: Modelling, Epistemic Authority, and Institutional Harm in AI Systems

Valeriano Grasso

Abstract

The dual-use problem of artificial intelligence is often treated as arising downstream issue arising at the point of deployment. We explore the hypothesis that dual-use is rooted upstream, in modelling and training practices that confer epistemic authority on AI systems before use. Focusing on design choices, data selection, and validation processes, we show how institutional reliance on AI systems displaces responsibility from human and organizational collectives onto the models themselves. With brief reference to cases, the paper reframes dual-use as a problem of epistemic legitimation, with direct relevance for accountability and peace-oriented AI research.

I. Introduction

The dual-use problem of artificial intelligence is commonly framed as a matter of deployment: models are developed for benign purposes and only later misused or adapted to harmful applications. For our purposes, 'dual-use' is understood not merely as contingent misuse, but as the possibility that the same epistemic artifacts are legitimately mobilized across heterogeneous and potentially conflicting institutional contexts. We propose that dual-use risks are already shaped upstream. Dual use does not begin at the point of use, but upstream, in the modelling and training practices that already embed orientations about what counts as relevant, reliable, and effective knowledge. This does not imply that downstream uses lack agency; rather, it suggests that such agency operates within epistemic constraints already stabilized upstream, as the legitimacy conferred during modelling and training conditions which uses are perceived as reasonable, justified, or institutionally acceptable.

Although AI systems lack consciousness or agency, they are increasingly treated as authoritative sources within institutional and public decision-making. This authority is not a property of the systems themselves, but the result of design choices, training regimes, validation practices, and institutional processes that stabilize models as epistemically trustworthy. When systems are prematurely or uncritically legitimized, responsibility for their outputs is transferred from the distributed human and institutional collectives that produced them onto the model itself. This transfer is visible in well-documented cases such as COMPAS and RoboDebt, where algorithmic systems were treated as authoritative decision-support infrastructures despite persistent contestation of their modelling assumptions However, these are not the only contexts in which epistemic authority is established upstream and subsequently applied in high-stakes settings. Beyond these well-studied cases, other institutional deployments illustrate how epistemic authority can be pre-established and applied in high-stakes settings. For example, contemporary uses of AI-based risk assessment systems in European border control — as documented in recent interdisciplinary work on automated decision-making at EU borders (Yang et al. 2024) — provide a paradigmatic case illustrating how these AI systems are institutionally positioned as authoritative, without assessing responsibility practices directly.

In such systems, epistemic authority is given well in advance through design choices, training regimes, and institutional validation processes that render model outputs trustworthy by default. What is at stake here is not merely the risk of error or bias, but a more subtle dynamic of legitimation and responsibility displacement. Operating in politically sensitive and quasi-militarized contexts,

algorithmic risk assessments are routinely framed as neutral inputs rather than as situated epistemic judgments shaped by prior assumptions embedded in data and model design.

Responsibility is thereby redistributed: not eliminated but displaced from the distributed human and institutional collectives involved in the design, training, and authorization of these systems onto the outputs themselves. Crucially, this dynamic is reinforced by opacity, both with respect to how specific decisions are produced and to how training data, validation criteria, and epistemic assumptions are constituted upstream. The authority conferred in advance is thus combined with limited contestability downstream, producing a situation in which trust in AI systems is institutionally and epistemically reinforced. Thus, dual use does not primarily arise from exceptional misuse, but from the prior stabilization of epistemic authority under conditions of opacity, enabling the same models to circulate across administrative, security, and enforcement contexts without fresh scrutiny.

By reframing dual-use as a structural feature of modelling rather than an exceptional outcome of misuse, we shift attention upstream. This hypothesis does not deny downstream agency or contingency but suggests that upstream epistemic legitimation limits the range of uses that are made plausible, legitimate, or institutionally acceptable. The central claim is that accountability in AI research cannot be adequately addressed by focusing on deployment alone, but calls for us to examine how epistemic authority is produced, stabilized, and institutionalized during modelling and training.

II. Modelling, Training, and the Production of Epistemic Authority

Following work in STS and philosophy of science, modelling can be understood not as a neutral technical operation, but as an epistemic practice that stabilizes what counts as relevant and authoritative knowledge (Latour, 2005; Star, 1999; Leonelli, 2016). When AI systems are granted epistemic authority within institutional decision-making, the downstream effects make upstream legitimation a central concern for peace-oriented AI research. We do not claim a linear causal chain but distinguish analytically among various sites where epistemic authority is progressively stabilized. Rather, we propose a structured conceptual exploration organized around three interrelated hypotheses concerning how epistemic authority in AI systems is produced and stabilized upstream. First, we examine how modelling and design choices pre-configure epistemic salience by deciding which features, categories, and forms of uncertainty are regarded as salient or irrelevant. Second, we explore how training regimes and data practices stabilize this epistemic authority by embedding prior assumptions in apparently neutral performance metrics. Third, we consider how validation procedures and institutional uptake extend this authority across contexts, enabling AI systems to circulate as trustworthy infrastructures beyond their original domains of development. Taken together, these hypotheses function as an analytical scaffolding for tracing how dual-use risks emerge as structural features of contemporary AI research and deployment.

Discussions of dual-use in AI research often focus on downstream applications, yet the production of epistemic authority begins much earlier, within the training phase of modelling. The inclusion of datasets in training already functions as a form of epistemic legitimation, defining what counts as admissible evidence. Training is where relevance and acceptable forms of knowledge are materially encoded. Through choices concerning data selection, labelling, loss functions, and evaluation, training orients what a system can meaningfully ‘respond to’ long before deployment. From an STS perspective, this phase functions as an epistemic bottleneck (Kelly, 2025), as far as training stabilizes model behaviour in ways that are difficult to revise once outputs are publicly and institutionally recognized as reliable. Once training stabilizes a model’s behaviours, its outputs are treated as expressions of technical reliability, while the socio-technical processes that produced them recede

into the background. This dynamic is reinforced by broader public imaginaries of AI. Models cannot be legitimized only within scientific/institutional communities. Indeed, the public sphere increasingly tends to treat algorithmic outputs as authoritative by virtue of their mediation through technical systems. As with other forms of screen-based credibility, AI outputs often confer legitimacy independently of their conditions of production. At the level of everyday epistemic validation, AI systems are commonly treated as if they possessed a form of moral or epistemic agency, as unified encyclopaedic sources of knowledge, with little awareness of the heterogeneity of their training regimes and the situated collectives that produce them. This everyday legitimation encourages users to treat LLMs as broadly authoritative advisors, extending epistemic trust well beyond expert contexts. Questions of responsibility are put on hold by the presumption that technical systems already operate within acceptable normative boundaries.

Researchers are often aware of these dynamics but structurally motivated to bracket them. Publication norms, benchmark-driven evaluation, and the pressure to demonstrate performance discourage impede reflection on the epistemic and social implications of training choices. This is a matter of institutional rigidity: certain critiques struggle to earn legitimacy unless they are already established within the field. As a result, accountability is frequently sought at the level of the system itself. Institutions – or the groups that finance and/or adopt AI technologies – often attempt to symbolically delegate responsibility to models treated as quasi-autonomous agents, despite the absence of any meaningful sense in which such responsibility can be transferred. The insistence on locating responsibility in the system signals the difficulty of confronting the distributed human and institutional commitments embedded in training and design. We thus suggest that dual-use risks may be better understood as systematically enabled by early epistemic legitimation, rather than as merely exceptional outcomes of misuse.

III. Conclusion

We have argued that the dual-use problem of artificial intelligence cannot be adequately addressed by focusing on deployment alone. We hypothesis that dual-use risks may be systematically enabled by early processes of epistemic legitimation during modelling and training, rather than arising solely from exceptional cases of misuse. Training practices determine what counts as admissible evidence and reliable output, thereby shaping the space of possible uses. Understanding dual-use as structurally rooted in epistemic legitimation shifts attention upstream. This understanding aligns with recent analyses of dual-use risks in AI research that emphasize upstream epistemic and institutional conditions rather than downstream misuse alone (Brenneis, 2025). While Brenneis frames dual-use as a governance challenge requiring upstream attention, Kelly's notion of epistemic bottlenecks helps explain how such risks become stabilized at the level of training and validation. Seen from this perspective, attempts to delegate accountability to AI systems are conceptually misguided. Responsibility cannot meaningfully reside in models treated as quasi-autonomous agents alone but remains distributed across the human and institutional arrangements that authorize them as epistemically trustworthy. The recurring search for an artificial bearer of responsibility signals not a solution, but a displacement: a way of avoiding confrontation with the commitments embedded upstream, at the level of training, design, and validation.

LLM Usage Statement.

I used a LLM for limited editorial assistance, including language/translation refinement and synthesis to follow call guideline. All conceptual development, argumentation, and references were produced and verified by the authors, who take full responsibility for the content of this paper.

Bibliografia

- Brenneis, A. (2025). *Assessing dual use risks in AI research: Necessity, challenges and mitigation strategies*. Research Ethics, 21(2), 302–330.
- Kelly, M. (2025). *Situated epistemic infrastructures: A diagnostic framework for post-coherence knowledge*. arXiv preprint arXiv:2508.04995.
- Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.
- Leonelli, S. (2016). *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.
- Star, S. L. (1999). "The Ethnography of Infrastructure." *American Behavioral Scientist*, 43(3), 377–391.
- Yang, Y., Zuiderveen Borgesius, F., Beckers, P., & Brouwer, E. (2024). *Automated decision-making and artificial intelligence at European borders and their risks for human rights* arXiv preprint arXiv:2410.17278.