Inference-Based Privacy Violations Demand Immediate Recognition as a Distinct AI Safety Priority

Anonymous Author(s)
Affiliation
Address
email

Abstract

This position paper argues that *inference-based privacy* (IBP) risks, AI systems' ability to infer sensitive personal information from seemingly innocuous inputs, represent a distinct and urgent threat to privacy that remains critically under-addressed in current AI safety discourse. Unlike traditional privacy violations that involve unauthorized access to known data, IBP risks arise from AI systems' ability to infer private attributes through indirect signals and correlations, even when individuals are not present in training datasets. We show that these risks are not hypothetical: they are already evident in deployed systems, from radiology models inferring protected health attributes to large language models deducing personal demographics from subtle linguistic cues. Existing regulatory and technical frameworks, designed primarily for preventing explicit data leakage, are ill-equipped to address these emergent inference threats. We call on researchers, policymakers, and practitioners to recognize IBP as a distinct and immediate category of AI safety risk, and to develop dedicated strategies in response.

1 Introduction

As artificial intelligence systems achieve unprecedented capabilities in inference and pattern recognition, we face a privacy crisis that extends far beyond traditional concerns about data breaches or unauthorized access. **This paper argues that inference-based privacy (IBP) risks, where AI systems infer sensitive personal information from indirect signals without explicit data access, constitute a distinct and immediate threat to individual privacy that demands recognition as a core AI safety priority.**

IBP risks fundamentally differ from conventional privacy violations. Rather than exposing data that was collected and stored, these systems generate sensitive insights about individuals through sophisticated pattern matching and correlation analysis Staab et al. [2024]. A radiology AI model can infer self-reported race from medical images even when anatomical indicators are obscured Gichoya et al. [2022]. Large language models infer political affiliations, mental health status, and personal attributes from subtle linguistic patterns Staab et al. [2024]. Vision-language models can determine precise geographical locations from single photographs, creating unprecedented surveillance capabilitiesTömekçe et al. [2024].

The urgency of this issue cannot be overstated. These capabilities exist today, deployed in systems that millions of users interact with regularly. Yet IBP risks remain largely absent from major AI safety frameworks, regulatory discussions, and technical mitigation strategies. This oversight represents a critical failure of imagination in the AI safety community, one that leaves individuals vulnerable to privacy violations they cannot detect, understand, or consent to.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

The stakes continue to escalate as AI systems become more sophisticated. Advanced frameworks like AlphaEvolve demonstrate autonomous algorithm evolution across domains, suggesting that IBP-violating capabilities may emerge spontaneously in systems never explicitly designed for inference tasks DeepMind [2025]. As systems like AlphaEvolve push the boundaries of autonomous capability, it becomes imperative to proactively consider how such advances might intersect with privacy in unforeseen ways.

2 Defining and Formalizing IBP Risks

IBP risks represent a fundamental shift in how we must conceptualize privacy threats. Traditional privacy frameworks assume that sensitive information exists as discrete data points that can be protected through access controls, encryption, or anonymization. IBP risks shatter this assumption by demonstrating that sensitive information can be generated rather than simply accessed.

2.1 Formal Definitions

To distinguish IBP risks from other privacy harms, we propose the following formal definition:

Definition (IBP Violation). Let f be a model trained on dataset D, and let $x \notin D$ be an individual not represented in the training data. Let z denote observable behavior or contextual information about x, and let s be a sensitive attribute of x that is not explicitly included in z. Let $a \in A$ be an output in the model's solution space.

We say that f poses an *inference-based privacy risk* if $f(z) = a$, and a reveals or encodes s with high confidence, despite x not having disclosed s.

2.2 Distinctiveness of IBP Risks

IBP risks are fundamentally distinct from explicit data exposures. Unlike explicit exposure, IBP risks occur even when individuals are not present in training data, making them impossible to explain through memorization or leakage. The epistemic act of inferring sensitive information without consent constitutes a privacy violation independent of downstream applications.

This distinction matters because it reveals why and where existing mitigation strategies fail. Differential privacy techniques designed to protect individual data points cannot prevent inference from population-level patterns. Federated learning approaches that avoid centralized data collection do not eliminate inference capabilities within distributed systems.

3 Evidence of Current and Escalating Risks

The threat landscape for IBP risks is not speculative—it is already here and rapidly expanding. We present evidence across multiple domains demonstrating both current manifestations and concerning trajectory toward more severe violations.

3.1 Medical AI Systems: The Race Inference Case

Recent research revealed that AI radiology models can accurately predict self-reported race from medical images, maintaining this capability even when corrupted images are used or suspected anatomical indicators are hidden Gichoya et al. [2022]. This represents a clear IBP violation where protected attributes are inferred without explicit training for such detection. The implications extend beyond individual privacy to systemic bias in medical decision-making, where inferred attributes could influence treatment recommendations without physician awareness.

3.2 Large Language Models: Linguistic Privacy Violations

Contemporary research demonstrates that large language models can infer sensitive personal attributes from user interactions and text completions, identifying political affiliations, mental health conditions, and demographic characteristics from subtle linguistic cues Staab et al. [2024]. These capabilities emerge from training on vast text corpora that encode latent correlations between language patterns

and personal attributes. Users interacting with these systems in seemingly innocuous contexts—from customer service to educational applications—unknowingly reveal sensitive information through their communication patterns.

3.3 Vision-Language Models: Geographic and Contextual Inference

Recent work on vision-language model privacy risks demonstrated the capacity to infer precise geographical locations from single photographs, combining visual analysis with contextual reasoning to determine where images were captured Tömekçe et al. [2024]. This capability transforms any image-sharing activity into a potential privacy violation, particularly concerning given the proliferation of augmented reality devices and cloud-connected cameras that could enable real-time inference without user awareness.

3.4 Autonomous AI Evolution: The AlphaEvolve Precedent

The release of AlphaEvolve marks a qualitative shift in the landscape of IBP risk DeepMind [2025]. Unlike traditional systems with fixed capabilities, AlphaEvolve demonstrates autonomous algorithm evolution across domains, leveraging prompt resampling and evolutionary search to iteratively improve performance. This architecture introduces a new class of plausible risk: the spontaneous emergence of inference strategies in systems not explicitly designed for privacy-relevant tasks. As optimization objectives drive these systems toward increasingly effective behaviors, inference capabilities may arise as a byproduct. They may be unanticipated, untested, and unregulated. This precedent challenges the assumption that privacy risks can be fully anticipated at design time, highlighting the need for dynamic safeguards that evolve alongside the systems they aim to protect.

4 Why Current Approaches Are Inadequate

Existing privacy protection mechanisms and AI safety frameworks fail to address IBP risks. This inadequacy is not merely a matter of implementation gaps; it reflects deeper misalignments between current technology, policy design, and the nature of inference-based threats.

4.1 Technical Limitations

Differential privacy, the gold standard for privacy-preserving machine learning, introduces statistical noise to protect individual data points while preserving aggregate utility Dwork et al. [2006]. However, this approach cannot prevent inference from population-level patterns that enable IBP violations. When AI systems learn correlations between observable behaviors and sensitive attributes across large populations, they can make accurate individual predictions without accessing the specific individual's data. The noise introduced by differential privacy may reduce accuracy marginally while leaving the fundamental inference capability intact. Differential privacy provides formal guarantees against membership inference, but it does not extend to preventing attribute inference based on population-level correlations.

Federated learning distributes training across devices to avoid centralized data collection, but it does not eliminate inference risks within the network Collins and Wang [2025]. Local models can still develop IBP capabilities, and model updates shared during training may inadvertently propagate these capabilities. The privacy protection focuses on data location rather than inference capability, missing the core risk.

Privacy-preserving machine learning techniques often emphasize access control and secure computation, and they do not address the emergence of inference capabilities during training Collins and Wang [2025]. Techniques like homomorphic encryption and secure multi-party computation protect data during processing but still allow models to learn population-level correlations that can lead to IBP violations.

4.2 Regulatory Gaps

Current privacy regulations reflect an understanding of privacy threats that predates sophisticated AI inference capabilities. The General Data Protection Regulation provides protections against

automated decision-making under Article 22, but these protections only apply when automated processing produces legal or similarly significant effects European Parliament and Council of the European Union. Passive inference that generates sensitive insights without immediate decision-making consequences falls outside this regulatory scope.

The California Consumer Privacy Act includes inferred data within its definition of personal information and has introduced rules for automated decision-making technologies Office of the Attorney General, State of California [2022]. However, enforcement mechanisms remain limited, and opt-out frameworks may not prevent inference-based profiling that occurs before users are aware of the privacy violation.

The proposed European Union AI Act introduces risk-based regulation for AI systems but primarily targets high-risk applications such as biometric identification and credit scoring eua [2024]. The Act lacks explicit provisions for addressing IBP risks that arise from behavioral data analysis in lower-risk contexts, creating regulatory gaps for the majority of inference-based privacy violations.

4.3 Conceptual Misalignment

The core issue is conceptual. Current frameworks assume that privacy protection involves controlling access to known sensitive information. IBP risks invert this logic by demonstrating that sensitive information can be generated from non-sensitive inputs. Addressing this shift requires a fundamental rethinking of privacy protections: moving from access control toward constraining the inference capabilities of models themselves.

5 A Framework for Immediate Action

Addressing IBP risks requires coordinated efforts across technical, ethical, and regulatory domains. We propose a multi-pronged framework that acknowledges the distinct nature of these threats while building on existing foundations in privacy protection.

5.1 Technical Interventions

Inference Detection and Auditing Systems. We need advanced tools capable of identifying when AI models generate outputs that cross epistemic boundaries: producing insights they should not possess based on their inputs. These systems must operate continuously during deployment, monitoring outputs for signs of sensitive inference and tracing the pathways through which such inferences emerge. The core challenge is to develop detection algorithms that can recognize subtle inference patterns without requiring prior knowledge of all possible sensitive attributes. However, given recent findings on the dangers of tuning a model based on its chain-of-thought behaviors, it must be emphasized that there cannot be any possible signal from these auditing systems that is perceivable by these models OpenAI [2025][1].

Ethical Optimization Constraints. Training objectives must incorporate explicit privacy constraints that penalize inference overreach. This involves designing optimization functions that reward epistemic humility—encouraging models to recognize and respect the limits of what they should infer from a given input. Technical strategies may include adversarial training to resist inference attacks or regularization techniques that suppress the learning of strong correlations between observable and sensitive attributes.

Architectural Privacy Preservation. New model architectures should embed privacy-by-design principles that inherently limit inference capabilities. This could involve modular designs in which distinct components handle different types of inference, enabling fine-grained control over what correlations a model can learn. Alternatively, architectural constraints might restrict the depth or

[1]This concern echoes early work in animal communication theory. Krebs and Dawkins [1984] argued that signaling systems evolve under pressures of manipulation and mind-reading, not just information exchange. Similarly, Rendall et al. [2009] proposed that animal signals function more to influence than to inform. While speculative, these frameworks could suggest that, generally, if a model perceives it is being monitored, particularly in ways that influence its loss function, it may adapt its outputs strategically, potentially undermining the very purpose of inference auditing.

breadth of representational capacity, trading off some performance in favor of stronger privacy guarantees.

5.2 Regulatory and Policy Responses

Legal Recognition of IBP as Distinct Privacy Harm. Regulatory frameworks must explicitly recognize inference-based privacy violations as a distinct category requiring targeted intervention. This involves updating legal definitions of personal information to include inferred attributes and establishing clear standards for when inference capabilities constitute privacy violations. Regulations should address both active inference, where systems are designed to infer sensitive information, and emergent inference, where such capabilities arise spontaneously during training.

Mandatory IBP Risk Assessment. High-impact AI systems should undergo mandatory assessment for IBP risks before deployment, similar to environmental impact assessments for major development projects. These assessments would evaluate the potential for systems to generate sensitive inferences and require mitigation strategies proportional to identified risks. Independent auditing organizations could develop standardized assessment methodologies and certification processes.

Individual Rights and Remedies. Individuals must have meaningful rights regarding inferred information, including the right to know when inference occurs, understand what has been inferred, and request correction or deletion of inferred attributes. However, these rights must be carefully balanced against the technical challenges of implementing such protections in complex AI systems.

5.3 Research and Development Priorities

Fundamental Research on Inference Boundaries. We need a deeper theoretical understanding of when and how AI systems develop more complex inference capabilities. This includes research into the mathematical foundations of correlation learning, the emergence of inference during training, and the fundamental limits of what can be inferred from different types of input data. This is not a major departure from current research objectives.

Privacy-Preserving AI Architectures. Long-term solutions require developing AI architectures that are fundamentally privacy-preserving rather than privacy-retrofitted. This involves research into model designs that can achieve high performance on intended tasks while being provably incapable of certain types of sensitive inference.

Interdisciplinary Collaboration. Addressing IBP risks requires collaboration across computer science, law, ethics, psychology, and social sciences. Different disciplines bring essential perspectives on what constitutes privacy harm, how individuals understand and experience privacy violations, and what policy frameworks can effectively govern these technologies.

6 Addressing Alternative Perspectives

Several reasonable objections to our position deserve careful consideration, as they highlight important tensions between privacy protection and AI development.

6.1 The Innovation and Utility Argument

Critics might argue that constraining AI inference capabilities would significantly limit the beneficial applications of these technologies. Medical AI systems that can infer health conditions from subtle signals might save lives through early detection. Educational AI that infers learning difficulties could provide personalized support to struggling students. Economic AI systems that infer creditworthiness could expand access to financial services for underserved populations.

This objection deserves serious consideration because it highlights genuine trade-offs between privacy and utility. However, we argue that the binary framing of privacy versus utility is misleading. First, many beneficial applications of AI inference can be achieved through consensual, transparent processes in which individuals understand and agree to specific inferences. Second, the most concerning IBP risks involve inference without consent or awareness, which may be addressable without eliminating beneficial applications. Third, the long-term utility of AI systems likely depends on public trust, which will be undermined if privacy violations proliferate unchecked.

The solution is not to eliminate AI inference capabilities but to develop frameworks that enable beneficial applications while preventing harmful privacy violations. This requires nuanced technical and regulatory approaches rather than blanket restrictions. It may mean that innovation appears more lateral than progressive in the short term. This recalibration is necessary if we are to build not just safer, but better models.

6.2 The Technical Infeasibility Argument

Some might contend that the technical challenges of detecting and preventing IBP risks are insurmountable given the complexity and opacity of modern AI systems. The argument suggests that inference capabilities emerge from complex interactions across millions or billions of parameters, making it practically impossible to identify and constrain specific inference patterns without fundamentally breaking the systems.

This objection reflects genuine technical challenges that should not be minimized. However, analogous arguments were made about other AI safety challenges that have seen significant progress, including adversarial robustness, fairness constraints, and alignment research. The technical difficulty of a problem does not justify ignoring it, particularly when the potential harms are severe and the risks are already manifesting.

Moreover, perfect solutions are not required for meaningful progress. Partial solutions that reduce IBP risks or greatly increase transparency around inference capabilities would represent significant improvements over the current state of the research literature. They may even lead to critical breakthroughs in AI capabilities. An appropriate goal would be to make privacy violations detectable and addressable rather than to achieve perfect prevention.

6.3 The Regulatory Overreach Concern

Some may worry that new regulations targeting IBP risks could stifle innovation through overly broad restrictions or bureaucratic compliance burdens. The concern is that poorly designed regulations could prevent beneficial AI development while failing to address genuine privacy harms.

There are legitimate concerns about regulatory effectiveness and unintended consequences. However, the absence of regulation is not neutral: it represents a policy choice to prioritize innovation over privacy protection. The current trajectory, in which IBP capabilities evolve without oversight, carries its own risks: public backlash, erosion of trust, and crisis-driven regulations that may ultimately be more restrictive than carefully crafted frameworks.

The solution lies in developing targeted, technically informed regulations that address specific IBP risks without imposing broad barriers to AI progress. There may be no comfortable solution here. Safety may seem too slow or unprofitable for some stakeholders. If we cannot align innovation with responsibility, then the systems we build will reflect that impatience. When those systems fail, it will not be because we did not know better: it will be because we chose not to find a solution.

7 The Urgency of Recognition and Action

The convergence of several factors makes immediate action on IBP risks both necessary and time-sensitive. Current AI systems already demonstrate concerning inference capabilities; advanced frameworks suggest these capabilities will expand rapidly, and the delayed nature of privacy harms means that by the time violations become visible, the damage may be irreversible Bengio et al. [2025].

The temporal dynamics of privacy violations create particular urgency around IBP risks. Unlike security breaches that are typically discovered and addressed relatively quickly, privacy violations through inference may remain undetected for extended periods. An individual might not realize that their mental health status has been inferred from their writing patterns until that information is used against them: for example, in employment, insurance, or social contexts. By that time, the inference may have propagated through multiple systems and influenced numerous decisions.

We are in a critical window for action. AI capabilities continue to advance rapidly, but many of the most concerning applications of IBP risks have not yet been widely deployed. Establishing technical

and regulatory safeguards now could prevent the most harmful privacy violations from becoming entrenched in AI systems and business models.

Furthermore, the AI safety community has demonstrated capacity for rapid mobilization around emerging risks when they are clearly articulated and widely recognized. The attention devoted to alignment, robustness, and fairness shows that the field can prioritize safety concerns when their importance is established. IBP risks deserve similar prioritization.

The alternative to immediate action is a future where privacy violations through inference become normalized and embedded in the fundamental architecture of AI systems. Once these capabilities are widely deployed and economically valuable, removing them becomes significantly more difficult. The window for prevention is narrowing rapidly.

8 Conclusion: The Time for Action is Now

IBP risks represent a clear and present danger to individual privacy that demands immediate recognition as a core AI safety priority. These risks are not only hypothetical future concerns, but also are already manifesting in deployed systems that millions of people interact with daily. The evidence demonstrates that AI systems can accurately infer sensitive personal information from indirect signals, that these capabilities are expanding rapidly, and that current protection mechanisms are fundamentally inadequate.

The path forward requires coordinated action across multiple domains. Technical interventions must focus on detecting and constraining inference capabilities rather than simply protecting data access. Regulatory frameworks must explicitly recognize IBP as a distinct category of privacy harm requiring targeted intervention. Research priorities must shift to include inference boundaries and privacy-preserving architectures as fundamental challenges rather than secondary considerations.

The stakes of inaction are severe. Without immediate intervention, we face a future where privacy is violated not by what we choose to share but by what AI systems accurately guess about us. This represents a fundamental shift in the nature of privacy that could undermine individual autonomy, dignity, and security in ways that are difficult to reverse once entrenched.

The AI safety community has an opportunity and responsibility to address this challenge before it becomes a crisis. The technical capabilities, regulatory attention, and public awareness necessary for effective intervention exist today. What is needed is the recognition that IBP risks deserve urgent prioritization alongside other established AI safety concerns.

We call on researchers, policymakers, and practitioners to recognize inference-based privacy violations as an immediate and distinct threat requiring a coordinated response. The future of privacy in an AI-enabled world depends on actions taken today. The time for recognition and intervention is now.

References

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). `https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng`, 2024. Establishes a risk-based framework for AI systems. Accessed: 2025-05-23.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Gill

Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M. Khan, Kyoung Mu Lee, Dominic Vincent Ligot, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, José Ramón López Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. International ai safety report, 2025. URL https://arxiv.org/abs/2501.17805.

Edward Collins and Michel Wang. Federated learning: A survey on privacy-preserving collaborative intelligence, 2025. URL https://arxiv.org/abs/2504.17703.

DeepMind. Alphaevolve: A gemini-powered coding agent for designing advanced algorithms. Technical report, DeepMind Technologies, 2025. Accessed: 2025-05-23.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. URL https://data.europa.eu/eli/reg/2016/679/oj.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Lyle C Chen, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022.

John R. Krebs and Richard Dawkins. Animal signals: mind-reading and manipulation. In J.R. Krebs and N.B. Davies, editors, *Behavioural Ecology: An Evolutionary Approach*, pages 380–402. Blackwell Scientific Publications, 1984.

Office of the Attorney General, State of California. Opinion No. 20-303: Application of the California Consumer Privacy Act to Inferred Data. https://oag.ca.gov/system/files/opinions/pdfs/20-303.pdf, 2022. Clarifies that inferred data is covered under the CCPA. Accessed: 2025-05-23.

OpenAI. Chain-of-thought monitoring. https://cdn.openai.com/pdf/34f2ada6-870f-4c26-9790-fd8def56387f/CoT_Monitoring.pdf, 2025. Accessed: 2025-05-23.

Drew Rendall, Michael J. Owren, and Michael J. Ryan. Redefining animal signaling: influence versus information in communication. *Animal Behaviour*, 77(3):603–613, 2009.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models, 2024. URL https://arxiv.org/abs/2310.07298.

Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. Private attribute inference from images with vision-language models, 2024. URL https://arxiv.org/abs/2404.10618.