Software Cannot Read the Room

Understanding the Limits of Technology for Efficient Digital Policy

Vicente Antonio Arias Gonzalez, Associate, The Global Initiative for Digital Empowerment

Policy Brief

Keywords:
digital governance, technology's limits, AI policy, multistakeholder collaboration

INTRODUCTION

The misalignment between policy objectives and technology's capabilities hinders human-centered digital transformation, especially as generative artificial intelligence (AI) software is increasingly adopted across industries, governments, and societies. When policymakers overlook technology's inherent limitations, they risk designing governance structures that are impractical, counterproductive, or misaligned with technical realities.

Policies that fail to align with technology's technical realities – e.g., its design, limitations, and practical applications – cannot address the historical shortcomings of existing governance regimes or the complexities of emerging challenges.

This policy brief examines four case studies, highlighting the risks of misaligning policy and technological constraints. The cases focus on the limitations of machine learning, data poisoning, synthetic data, and AI automation. The brief concludes by emphasizing that, while multistakeholder approaches offer the most effective path to aligning governance objectives with technical constraints, their success ultimately depends on a genuine understanding of technology's limitations.

MACHINE (UN)LEARNING AND THE RIGHT TO BE FORGOTTEN

Due to the Brussels effect (Bradford, 2020), most recent data protection and privacy laws, worldwide, assume a level of technical feasibility that does not always align with real-world constraints, particularly in the context of AI systems. This misalignment complicates enforcement, increases compliance burdens on busi-

nesses and regulators, weakens intended protections for individuals, and overwhelms courts with excessive litigation.

Two key areas where this tension is evident are the interplay between the right to be forgotten, machine learning techniques, and (as shown below) lawful data processing regarding the limitations of synthetic data.

Lawful data processing grounds form the foundation of data protection regulations, determining when and how personal data can be collected, used, or otherwise processed. The right to be forgotten allows individuals to request the deletion of data when, for example, it is no longer necessary for its original purpose or when consent is withdrawn. However, while this principle is enforceable in structured databases, its implementation becomes significantly more complex in deep-learning models. Unlike databases, where specific information can be deleted, these models encode data in complex, non-interpretable ways, making complete removal nearly impossible.

Still, leading experts and policymakers have proposed "machine unlearning" as a key solution for implementing the right to be forgotten (Hine et al., 2024). Machine unlearning refers to the process of selectively removing specific training data points—and their influence—from an already trained model. The goal is for the updated model to behave as if it had never been trained on those data points at all (Xu et al., 2024).

However, machine unlearning has inherent limitations, both in back-end (removing training data effects from models) and front-end considerations (suppressing specific content in model outputs), raising concerns about its effectiveness in enforcing the right to be forgotten (Cooper et al., 2024).

The gold standard – removing specific data points and the full retraining of the software – provides probabilistic rather than absolute guarantees of information removal (Liu et al., 2024). Even in these cases, foundation models may generate novel content resembling removed information or reintroduce it through indirect associations.

"Un-unlearning" (Shumailov et al., 2024) occurs when Large Language Models (LLMs) generalize tasks based on descriptions, even when those tasks were not explicitly included in the training data (Agarwal et al., 2024). This capability can lead to unpredictable model outputs, causing previously unlearned knowledge to resurface, inadvertently reintroducing data that were meant to be removed. This challenge goes beyond machine unlearning, as governance frameworks rarely regulate how users interact with models. It is still unclear whether LLMs process personal data (Hamburg Data Protection Authority, 2024).

A related challenge for the right to be forgotten is the misunderstanding of data-sharing practices in the context of free data flows. A court ruling in Norway highlighted this issue when a company shared personal data with third parties, spreading it to thousands of entities (Forbrukerrådet, 2020). This case illustrated how, once data are shared, they can quickly expand beyond the original controller's reach, making the enforcement of deletion rights increasingly complex, even without the shortcomings of emerging technologies.

»Unlike industry-specific models, which are trained on controlled inputs, foundation models ingest data from unverified sources, making them particularly vulnerable to poisoning attacks.«

POISONED DATA AND POISONED SOFTWARE

Data poisoning techniques manipulate AI software by altering training data at any stage of their lifecycle, compromising their integrity and reliability. These techniques are particularly harmful in contexts where accuracy is essential, such as election integrity, political discourse, financial advising, and healthcare. This is because even small, low-cost batches of poisoned data can destabilize entire datasets (Alber et al., 2024).

Industry-focused AI software primarily relies on high-quality datasets but may still incorporate information from the Internet, where even verified sources can contain outdated or misleading research. Foundation models, including LLMs, rely on vast, indiscriminately sourced data rather than curated, industry-specific datasets, increasing the likelihood of data poisoning. (e.g., Sadeghi & Blachez, 2025) Unlike industry-specific models, which are trained on controlled inputs, foundation models ingest data from unverified sources, making them particularly vulnerable to poisoning attacks (Carlini et al., 2024).

In some cases, techniques such as embedding misleading text or employing "tarpits" designed to confound AI scrapers may inadvertently introduce corrupted data into AI training sets, even though their primary purpose is to prevent a website from being exploited as a data farm (Belanger, 2025).

In general, an attacker cannot directly control how data are labeled, making it more challenging to manipulate AI model behavior through traditional means. Additionally, since none of the leading LLMs are open source (Open Source Initiative, 2025), attackers do not have access to the trained model either, theoretically limiting their ability to affect AI software (VanHoudnos et al., 2024). However, attackers have developed methods to indirectly influence AI models by circumventing data controls, exploiting retrieval systems, and injecting adversarial prompts through AI's software pipeline (He et al., 2025).

The challenge of unlearning poisoned inputs creates additional difficulties. Once manipulated data influence a model, isolating and removing their impact is often too burdensome, particularly for smaller AI developers or those outside the leading competitors. As a result, AI models become less reliable, their performance degrades, and decision making is compromised across various applications. This is especially problematic for automated fact checking, where poisoned content weakens the ability of models to distinguish between credible and misleading infor-

mation. Consequently, content moderation becomes less effective, more error-prone (Du et al., 2022), and (as shown below) increasingly burdensome for the workers responsible for verifying the accuracy of autonomous systems.

This issue compounds broader technological shortcomings, such as perpetuating historical biases, susceptibility to hallucinations, and the knowledge degradation and response distortions introduced by fine-tuning mechanisms (Ghosh et al., 2024).

SYNTHETIC DATA

Synthetic data has emerged as a potential solution for mitigating privacy risks while enabling AI development because it replaces personal data with artificially generated information. Advanced generative models learn patterns from these data to produce new, seemingly realistic non-personal data. This approach allows organizations to navigate privacy regulations by enhancing compliance or structuring data practices to avoid direct regulatory oversight while they continue to benefit from data analysis and product development, especially as leading AI companies race to build more powerful AI systems (Conroy et al., 2025).

Although generative models for creating synthetic data are considered state-of-the-art, their privacy benefits remain unpredictable. A key limitation is determining which features from the original data are retained in the synthetic dataset (Stadler et al., 2022). As a result, for example, personal attributes such as ethnicity or income may still be inferred, raising concerns about re-identification risks (Hittmeir et al., 2020).

To address these concerns, businesses and experts argue that differential privacy can help reduce re-identification risks by adding "noise" – small random changes – to the data, which obscures individual details while maintaining overall trends and patterns in the synthetic dataset (Kurakin et al., 2023). However, while combining synthetic data with differential privacy techniques could offer stronger safeguards than traditional anonymization, higher privacy settings can result in significant utility loss, making synthetic data impractical for many use cases. Stronger privacy protections often lead to wider deviations from the original data, which can reduce the dataset's accuracy. In that sense, synthetic data do not provide a better trade-off between privacy and utility than traditional anonymization techniques. They also lead to unpredictable utility loss and highly unpredictable privacy gain (Sarmin et al., 2024).

As a result, balancing privacy and data utility is especially challenging when companies compete for market dominance or operate with limited resources. History has shown that utility normally prevails over human-centered features. Occasionally, in the race to develop innovative AI systems and gain a competitive edge, leading organizations may prioritize synthetic data that closely mirror real datasets, even at the expense of stronger privacy safeguards.

In this context, regulatory frameworks may struggle to keep pace with the competing demands of privacy compliance and utility-driven innovation, risking a scenario where compliance requirements are met on paper but fail to deliver substantive privacy protections. In short, since synthetic data does not eliminate re-identification

risks or resolve the privacy–utility trade-off, it is not a reliable safeguard. Overreliance on it may lead to regulations that satisfy formal compliance yet fall short of genuinely protecting personal data and people's well-being.

THE AUTONOMATION FALLACY

Policy frameworks and industry standards for AI safety, fairness, and accountability often assume that AI software operates mostly autonomously, minimizing the role of human labor at every stage of its lifecycle (Crawford, 2021). As a result, discussions often center on algorithmic flaws, explainability, and ex-post oversight, while overlooking how the systemic exploitation of human labor undermines these norms and their effectiveness.

The high expectations of automation obscure the extent to which humans are forced to compensate for technology's limitations (Williams et al., 2022). Governance frameworks prioritize protections for workers whose jobs may be disrupted by AI rather than those whose labor actively supports its development (OECD, 2024), even when direct references are made to low- and middle-income economies (UNESCO, 2021).

»The high expectations of automation obscure the extent to which humans are forced to compensate for technology's limitations.«

This narrow focus leaves the people essential to AI's functioning unprotected, reinforcing the misconception of full automation and reducing the effectiveness of existing oversight mechanisms.

Workers train models in image recognition, speech processing, and content moderation, correct AI errors, and intervene when automated systems fail. However, because AI labor is often outsourced across borders, workers frequently annotate or label data that reflect contexts, objects, or phenomena with which they have little or no direct experience (Muldoon et al., 2024).

More importantly, basic workers' rights, including fair wages, work benefits, and job stability, remain unaddressed, leaving AI workers vulnerable to exploitative conditions (Shield the Future, 2025). Many AI workers endure extreme psychological strain, with post-traumatic stress disorder (PTSD), insomnia, and other stress-related disorders being common; however, mental health support is usually inaccessible. Moreover, most earn as little as US$2 per hour (Perrigo, 2023).

These conditions highlight a critical limitation in governance frameworks that assume AI systems function autonomously, overlooking the essential human labor required to sustain them. AI safety, fairness, and accountability cannot be ensured when the workers who sustain these systems endure exploitative conditions. Furthermore, even if existing governance frameworks acknowledged the human labor behind AI, they do little to protect these workers (Bengio et al., 2025).

For instance, The United Nations Educational, Scientific and Cultural Organization's (UNESCO's) Readiness Assessment Methodology (RAM) process serves as an assessment tool for evaluating national AI governance readiness. While it offers policy guidance on ethical, regulatory, and infrastructure challenges, it overlooks the precarious conditions that AI workers face and offers no concrete safeguards to prevent their exploitation. Ignoring this reality not only fails workers but also undermines AI itself.

As AI expands, so will its reliance on human oversight, intervention, and maintenance. Governance frameworks must acknowledge this reality and implement enforceable policies that protect both AI users and the workers essential to its functionality.

CONCLUSION

Policy solutions that ignore the edges of existing technology – where it sometimes overperforms and sometimes falls short – often become part of the problem, worsening the very issues they aim to resolve.

Due to its inclusive and consensus-based nature, regional and national leaders should adopt multistakeholder collaboration processes to develop future digital governance frameworks, bridging policy goals and technological constraints. However, true collaboration demands more than drafting and circulating documents among isolated experts; it requires ongoing, interdisciplinary engagement, with expertise being continuously shared, refined, and adapted as challenges evolve.

At the same time, it is crucial to recognize that the technological limitations underlying existing governance designs play a pivotal role in shaping policy outcomes. Technology's limitations extend beyond what AI cannot do, including its potential to exceed our expectations. Effective governance requires acknowledging these constraints, ensuring that AI is neither expected to do more than it can realistically achieve nor overlooked when it surpasses assumed boundaries.

Effective digital policymaking ultimately depends on maintaining a clear awareness of technology's limitations, which is better achieved through ongoing multistakeholder processes.

»Policy solutions that ignore the edges of existing technology – where it sometimes overperforms and sometimes falls short – often become part of the problem, worsening the very issues they aim to resolve.«

REFERENCES

Agarwal, R., Singh, A., Zhang, L.M., Bohnet, B., Chan, S., Zhang, B., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J.D., Chu, E., Behbahani, F.M., Faust, A., & Larochelle, H. (2024). *Many-shot in-context learning.* arXiv, abs/2404.11018. URL: https://arxiv.org/abs/2404.11018

Alber, D.A., Yang, Z., Alyakin, A., Yang, E., Rai, S., Valliani, A.A., Zhang, J., Rosenbaum, G.R., Amend-Thomas, A.K., Kurland, D.B., Kremer, C.M., Eremiev, A., Negash, B., Wiggan, D.D., Nakatsuka, M.A., Sangwon, K.L., Neifert, S.N., Khan, H.A., Save, A.V., Palla, A., Grin, E.A., Hedman, M., Nasir-Moin, M., Liu, X.C., Jiang, L.Y., Mankowski, M.A., Segev, D.L., Aphinyanaphongs, Y., Riina, H.A., Golfinos, J.G., Orringer, D.A., Kondziolka, D., & Oermann, E.K. (2025). *Medical large language models are vulnerable to data-poisoning attacks.* Nature Medicine, 31, 618–626. URL: https://doi.org/10.1038/s41591-024-03445-1

Belanger, A. (2025, January 28). *AI haters build tarpits to trap and trick AI scrapers that ignore robots.txt.* Ars Technica. URL: https://arstechnica.com/tech-policy/2025/01/ai-haters-build-tarpits-to-trap-and-trick-ai-scrapers-that-ignore-robots-txt/

Bengio, Y., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., Newman, J., Ng, K.Y., Okolo, C.T., Raji, D., Sastry, G., Seger, E., Skeadas, T., & South, T. (2025). *International Scientific Report on the Safety of Advanced AI.* URL: https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf

Bradford, A. (2020). *The Brussels Effect: How the European Union Rules the World.* Faculty Books, 232. Columbia Law School

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., & Tramèr, F. (2024). *Poisoning web-scale training datasets is practical.* In *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 407–425). IEEE.

Conroy, L., Fehres, A., & AI4Media. (2025). *Tech companies are turning to synthetic data to train AI models, but there's a hidden cost.* The Conversation. URL: https://theconversation.com/tech-companies-are-turning-to-synthetic-data-to-train-ai-models-but-theres-a-hidden-cost-246248

Cooper, A.F., Choquette-Choo, C.A., Bogen, M., Jagielski, M., Filippova, K., Liu, K.Z., Chouldechova, A., Hayes, J., Huang, Y., Mireshghallah, N., Shumailov, I., Triantafillou, E., Kairouz, P., Mitchell, N., Liang, P., Ho, D.E., Choi, Y., Koyejo, S., Delgado, F., Grimmelmann, J., Shmatikov, V., Sa, C.D., Barocas, S., Cyphert, A., Lemley, M.A., boyd, D., Vaughan, J.W., Brundage, M., Bau, D., Neel, S., Jacobs, A.Z., Terzis, A., Wallach, H., Papernot, N., & Lee, K. (2024). *Machine unlearning doesn't do what you think: Lessons for generative AI policy, research, and practice.* arXiv, abs/2412.06966. URL: https://arxiv.org/abs/2412.06966

Crawford, K. (2021). *The Atlas of AI: Power, politics, and the planetary costs of artificial intelligence.* Yale University Press. URL: https://doi.org/10.2307/j.ctv1ghv45t

Du, Y., Bosselut, A., & Manning, C. D. (2022). *Synthetic disinformation attacks on automated fact verification systems.* Proceedings of the AAAI Conference on Artificial Intelligence, 36(10), 10581–10589. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21302

Forbrukerrådet. (2020). *Out of control: How consumers are exploited by the online advertising industry.* URL: https://storage02.forbrukerradet.no/media/2020/01/2020-01-14-out-of-control-final-version.pdf

Ghosh, S., Evuru, C.K., Evuru, R., Kumar, S., Ramaneswaran, S., Aneja, D., Jin, Z., Duraiswami, R., & Manocha, D. (2024). *A closer look at the limitations of instruction tuning.* arXiv, abs/2402.05119. URL: https://arxiv.org/abs/2402.05119

He, P., Xing, Y., Xu, H., Xiang, Z., & Tang, J. (2025). *Multi-faceted studies on data poisoning can advance LLM development.* arXiv, abs/2502.14182. URL: https://arxiv.org/html/2502.14182v1

Hine, E., Novelli, C., Taddeo, M., & Floridi, L. (2024). *Supporting trustworthy AI through machine unlearning.* Science and Engineering Ethics, 30(5). URL: http://dx.doi.org/10.1007/s11948-024-00500-5

Hittmeir, M., Mayer, R., & Ekelhart, A. (2020). *A baseline for attribute disclosure risk in synthetic data.* In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy (CODASPY '20)* (pp. 133–143). Association for Computing Machinery. URL: https://doi.org/10.1145/3374664.3375722

Kurakin, A., Ponomareva, N., Syed, U., MacDermed, L., & Terzis, A. (2023). *Harnessing large-language models to generate private synthetic text.* arXiv, abs/2306.01684. URL: https://arxiv.org/abs/2306.01684

Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Xu, X., Yao, Y., Liu, C., Li, H., Varshney, K.R., Bansal, M., Koyejo, S., & Liu, Y. (2024). *Rethinking machine unlearning for large language models.* arXiv, abs/2402.08787. URL: https://arxiv.org/abs/2402.08787

Muldoon, J., Graham, M., & Cant, C. (2024). *Feeding the machine: The hidden human labour powering AI.* Canongate Books.

OECD. (2024). *OECD Employment Outlook 2024: The net-zero transition and the labour market.* OECD Publishing, Paris. URL: https://doi.org/10.1787/ac8b3538-en

Open Source Initiative. (2024). *The Open Source Definition (OSD).* URL: https://opensource.org/osd

Perrigo, B. (2023). *OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic.* TIME. URL: https://time.com/6247678/openai-chatgpt-kenya-workers/

Sadeghi, M., & Blachez, I. (2025). A well-funded Moscow-based global 'news' network has infected Western artificial intelligence tools worldwide with Russian propaganda. NewsGuard's Reality Check. URL: https://www.newsguardrealitycheck.com/p/a-well-funded-moscow-based-global

Sarmin, F.J., Sarkar, A.R., Wang, Y., & Mohammed, N. (2024). *Synthetic data: Revisiting the privacy-utility trade-off.* arXiv, abs/2407.07926. URL: https://arxiv.org/abs/2407.07926

Shield the Future (2025). *The Humans Behind Machines.* [Video]. YouTube. URL: https://youtu.be/2GF5TVcjmv4

Shumailov, I., Hayes, J., Triantafillou, E., Ortiz-Jiménez, G., Papernot, N., Jagielski, M., Yona, I., Howard, H., & Bagdasaryan, E. (2024). *UnUnlearning: Unlearning is not sufficient for content regulation in advanced generative AI.* arXiv, abs/2407.00106. URL: https://arxiv.org/abs/2407.00106

Stadler, T., Oprisanu, B., & Troncoso, C. (2022). *Synthetic data – Anonymisation groundhog day.* Proceedings of the 31st USENIX Security Symposium. URL: https://www.usenix.org/system/files/sec22-stadler.pdf

The Hamburg Commissioner for Data protection and freedom of information. (2024). *Diskussionspapier: Large language models und personenbezogene Daten.* URL: https://datenschutz-hamburg.de/fileadmin/user_uplo

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence.* URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455

VanHoudnos, N., Smith, C., Churilla, M., Lau, S. H., McIlvenny, L., & Touhill, G. (2024). *Counter AI: What is it and what can you do about it?.* URL: https://insights.sei.cmu.edu/documents/6016/Counter_AI_What_Is_It_and_What_Can_You_Do_About_It_Brochure_6_z6T338h_9MPLSLQ.pdf

Williams, A., Miceli, M., & Gebru, T. (2022). *The exploited labor behind artificial intelligence.* Noema Magazine. URL: https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/

Xu, J., Wu, Z., Wang, C., & Jia, X. (2024). *Machine unlearning: Solutions and challenges.* IEEE Transactions on Emerging Topics in Computational Intelligence. URL: https://arxiv.org/abs/2308.07061v2