Review

AI Goes Off the Grid: The Rise of Local AI Demands Rethinking AI Governance

Bahrad A. Sokhansanj *

Posted Date: 9 June 2025

doi: 10.20944/preprints202506.0680.v1

Keywords: artificial intelligence; AI policy; AI ethics; AI safety; digital governance; open-source AI; generative AI; local AI)

 doi:10.20944/preprints202506.0680.v1

Article

AI Goes Off the Grid: The Rise of Local AI Demands Rethinking AI Governance

Bahrad A. Sokhansanj [1,2]

[1] Department of Electrical & Computer Engineering, College of Engineering, Drexel University, Philadelphia, PA 19104; bahrad@bahradlaw.com
[2] Law Office of Bahrad Sokhansanj, Los Angeles, CA 90034

Abstract: The centralized AI governance paradigm is breaking down. While policymakers focus on regulating cloud-based systems that run on massive, power-hungry data centers operated by big companies like Google and OpenAI, a revolution in the AI ecosystem unfolds. Open-source AI models can now run on personal computers and devices, invisible to regulators and stripped of safety constraints. Recent software and hardware advances mean that the capabilities of local-scale AI models now lag just a few months behind those of state-of-the-art proprietary models. Local AI has profound benefits for privacy and autonomy. But local AI also fundamentally disrupts AI governance. Technical safeguards fail when users control the code, and regulatory frameworks collapse when deployment becomes invisible. In this paper, we review how decentralized, open-source local AI undermines both technical and policy-based AI governance mechanisms. We propose ways to reimagine AI governance for these new challenges through 1) novel approaches to technical safeguards, including content provenance, configurable safe runtime environments, and distributed project monitoring, with 2) policy innovations including polycentric governance, participatory community approaches, and tailored safe harbors for liability. These proposals aim to catalyze a broader dialogue on harnessing local AI's democratizing potential while managing its risks and reinforcing ethical accountability.

Keywords: artificial intelligence; AI policy; AI ethics; AI safety; digital governance; open-source AI; generative AI; local AI)

1. Introduction

The generative artificial intelligence (AI) revolution began in research labs but became a mass phenomenon in November 2022, when OpenAI released ChatGPT, a powerful large language model (LLM) delivered through an easy-to-use web-based chatbot [1]. This breakthrough represented a major shift from conventional machine learning's focus on prediction and classification towards AI systems designed to create novel content. The scope of generative AI extends beyond LLMs to encompass vision-language models (VLMs), audio synthesis systems, and image and video generation tools. The ability to generate and execute code using LLMs opens the door to semi-autonomous or autonomous "agent" systems capable of reasoning and problem-solving [2,3]. Autonomous agents can potentially even teach themselves new capabilities [4–6]. Generative AI has become so prevalent that the term "AI" has become synonymous with it in popular usage, even though conventional machine learning models have existed for decades. ("Generative AI" and "AI" are used interchangeably throughout this paper, consistent with most contemporary literature on LLMs and related models.)

Companies like OpenAI, Anthropic, and Google deploy increasingly sophisticated models that are accessed through the web and run on data centers containing massive GPU clusters [7]. Their mode of operation establishes clear points of control: corporate providers can monitor usage, enforce safety guardrails, and implement pricing structures that shape how these technologies are used. However, the landscape is now undergoing another transformation. Powerful open-source models have emerged that can run outside institutional providers' cloud-based services on local hardware. Figure 1 summarizes

 doi:10.20944/preprints202506.0680.v1

how this shift represents a fundamental change from provider-controlled, centralized infrastructure to consumer-controlled, decentralized deployment.

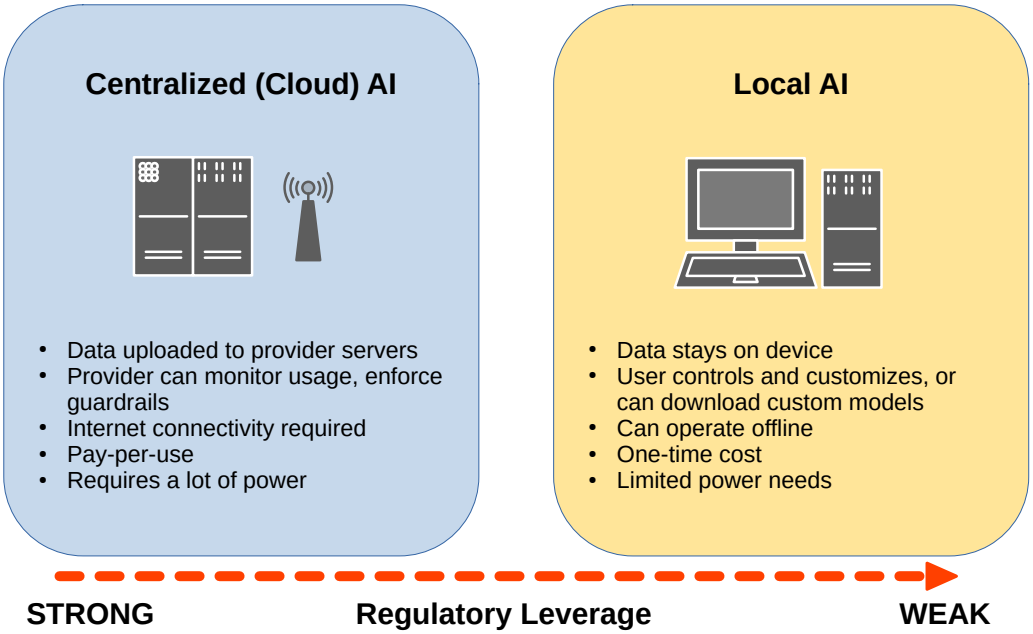

Figure 1. On the left, centralized (cloud-based) AI systems—such as OpenAI's ChatGPT, Google's Gemini, or Anthropic's Claude—route user data to provider-controlled GPU clusters. This allows for the enforcement of strong alignment guardrails, monitoring of inputs, and refusal of potentially harmful outputs. The infrastructure for cloud-based AI requires buying a lot of specialized hardware, maintaining a large physical footprint, and consuming a high level of electrical power, all of which allow governments to monitor it and enforce regulations. On the right, local AI runs directly on consumer hardware, enabling private, offline use of more limited but still highly capable models. These models can also be modified to remove internal safety constraints. Lacking large-scale infrastructure, local AI can operate invisibly, and, as a consequence, pose significant challenges for oversight.

The emergence of open-source AI fundamentally transforms who can access and modify advanced AI capabilities—and, by extension, who can potentially misuse them [8]. The implications for AI governance are profound, since access and use restrictions can no longer be enforced without centralized commercial providers like OpenAI or Google. Indeed, locally deployable open-source AI, or "local AI," presents even further challenges beyond those of open-source systems. When open-source models run on data centers, they generally require significant investment and a physical presence that facilitates oversight. By contrast, local generative AI can run on consumer hardware, including personal computers, laptops, and, as hardware improves, even smartphones—all without constant connection to cloud services or external servers. Local AI is much harder to regulate as a result.

While developments in open-source and local AI have been less broadly publicized than the introduction of new versions of ChatGPT, the "open source" AI paradigm was thrust into the spotlight in early 2025 with the release of DeepSeek-R1, an open-source model developed in China that provided performance comparable to large closed-source cloud-based models [9]. DeepSeek-R1 is a large model that has to run on multiple servers in a data center. However, as an open-source model, it can be hosted anywhere and not just on DeepSeek's own China-based servers. DeepSeek also released "distillations" of R1—versions of local models like Llama and Qwen fine-tuned on the output of the larger model.

DeepSeek's announcement heralded the potential of a simple distillation approach that can train on the output of big models to fine-tune local models. This approach allows developers to boost the performance of locally deployable models to levels that prove useful for many applications. And in some use cases, local model performance is now comparable to that of state-of-the-art proprietary systems [9]. This represents a critical turning point: The initial training of large-scale base models still

 doi:10.20944/preprints202506.0680.v1

requires massive data centers that could be theoretically regulated. However, after that training has taken place, big models can in turn be used to generate synthetic datasets that allow the power of proprietary foundational models to be transferred to create smaller, yet still highly performant models that can run locally. This paper explores the governance challenges of local AI in depth. Section 2 provides background on local AI, examining why users choose local deployment and highlighting potential high-risk applications across biosecurity, information integrity, and cybersecurity domains. Section 3 analyzes how local AI disrupts the conventional AI safety paradigm, showing how local deployment undermines technical safeguards and violates fundamental assumptions of current AI regulatory frameworks. Section 4 offers proposals that represent a starting point for rethinking governance in res vcponse to the emerging challenges of the local AI ecosystem: 1) developing novel approaches to technical safeguards, such as content provenance technologies, secure computation environments, and distributed monitoring systems; and 2) policy innovations including polycentric governance frameworks, community-driven participatory models, and safe harbor legal protections for responsible actors. These proposals are grounded in the urgent need for multi-layered AI governance capable of addressing a spectrum of implementations that range from massive data centers to personal computers located in homes and offices. They are intended to be a proactive starting point that anticipates the inevitable progress of software and hardware towards enabling a powerful local AI ecosystem.

2. Background

2.1. Why Go Local?

The first prominent open-source LLM that began to approach the capabilities of ChatGPT was Meta's Llama[10], followed soon thereafter by Mistral's models[11]. While Meta and Mistral provided access to their models through their own API endpoints, such as Mistral LeChat, they also provided open-source versions that could be deployed locally on computers with sufficient resources or self-hosted outside the reach of the model publishers. These models have been joined by powerful open-source models, including Google's Gemma 3 [12], Microsoft's Phi-3 [13], Alibaba's Qwen 3 [14], and the United Arab Emirates' Falcon [15].

Moreover, techniques such as quantization and caching have been developed and adopted, which reduce the memory load of LLM inference and training, allowing them to be effectively run on progressively more commodified, cheaper, and energy-efficient processors [16–21]. Another model architecture innovation that enables local LLMs is "Mixture-of-Experts," in which only a subset of the model's parameters (the "experts") are activated for any given input, with a gating mechanism determining which experts to use [22,23]. This architecture reduces computational requirements during inference, making larger models feasible on local hardware. When combined with memory management techniques in open-source implementations, this enables efficiently dividing model inference such that each time it runs, a fraction of the model needs to be loaded in GPU memory (VRAM, more expensive) and the rest of a big model in ordinary RAM (much cheaper), enabling much bigger models to run locally [14,24].

Figure 2 charts the rapid improvement of open-source models that can potentially be run on local machines, showing how it has paralleled the progression of large proprietary models that are considered the flagships of companies like OpenAI and Google. The ability to run models locally is set to advance even further through the emergence of "AI PCs," such as the Nvidia DIGITS system, which has a powerful GPU and on-board memory, allowing it to run larger-scale models [25]. Apple offers M4 chips, which can run AI models on battery-powered laptops [26]. AMD has also recently announced a "workstation-class" GPU that is aimed at developers and professionals using AI, rather than the gamers who have until now dominated the consumer GPU market [27].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 June 2025 doi:10.20944/preprints202506.0680.v1

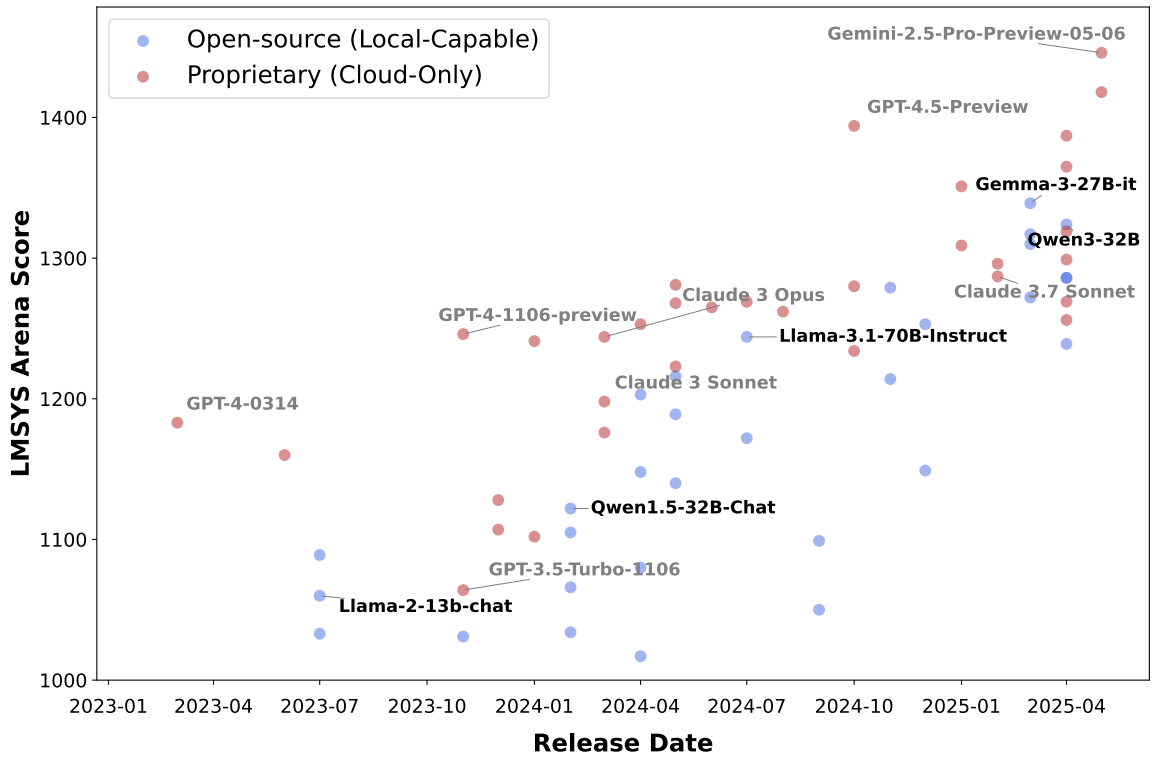

Figure 2. Recently, the performance of locally runnable open-source language AI models has improved in parallel with proprietary cloud-provided models. The scores shown here are Elo values from the LMSYS Chatbot Arena, a crowd-sourced, head-to-head evaluation platform that places closed and open models on a single, continuously updated scale [28]. While it is a less quantitatively rigorous benchmark than others, it provides a metric of the real-world usefulness of models. Blue points (with some exemplary model results labeled with black text) mark open-source models that can be executed on high-end consumer hardware; red points (labeled with grey text) mark proprietary systems from OpenAI, Google, and Anthropic released in the same period. These data points exemplify how, in less than two years, leading local models have moved from far-behind contenders to lagging only a few months behind the proprietary flagships.

Some skeptics may still question whether local AI will play a significant role in the overall AI ecosystem, given that proprietary cloud-provided systems are becoming more capable as well. However, there are compelling reasons to go local. As an initial matter, local AI can save on costs, since running models on a local device means avoiding paying fees to use cloud services through APIs and web applications. Any such cost savings, to be sure, can be somewhat negated by the need to buy more powerful computers to run more capable AI models. Even so, the aforementioned competition emerging among chipmakers should reduce costs over time. Local AI also avoids the need for an Internet connection, though that is also mitigated by the increasing prevalence of Internet connectivity even in remote areas and on airplanes. There are further deeper personal and social benefits to moving to a local, decentralized delivery of AI. These include control over user privacy, autonomy from large commercial providers, and greater customizability outside of a centralized platform.

First, the privacy-preserving aspect is particularly important in domains with strict confidentiality requirements, such as healthcare [29]. To ensure patient confidentiality, researchers have deployed local LLMs for anonymizing radiology reports[30]. and offline-capable chatbots for self-managing hypertension [31]. Other domains have similar confidentiality requirements, such as legal and financial services, and proprietary business operations [32]. For example, in law, a key obstacle to AI use is that when using cloud platforms, there is a risk that confidential attorney-client communications and work product are recorded by logging prompts and responses, risking potential security breaches or even loss of privilege in court proceedings [33,34]. These data privacy concerns for attorneys are obviated by self-hosted or local AI systems.

 doi:10.20944/preprints202506.0680.v1

Second, local deployment reduces dependency on AI providers who might otherwise raise prices, restrict access based on commercial considerations, impose usage terms, or even discontinue services. Locally deployed models are becoming increasingly viable on lower-cost systems, though with performance trade-offs as devices are less capable [35]. Additionally, local deployment can prevent the need to use foreign cloud providers, such as using China's DeepSeek models in the United States [35]. Cloud-based providers also implement restrictions on what their systems will discuss or assist with. LLMs will deny user requests with automated responses based on safety restrictions, but those can be drawn so broadly as to capture even legitimate uses, such as for education, political organizing, and politically sensitive topics [36]. Local AI can thus help counterbalance the concentration of power in a few dominant technology companies, and democratize access to advanced capabilities [37–39].

Third, local deployment offers greater autonomy. Local AI avoids dependence on platforms that may enforce ideological constraints, commercial gatekeeping, or compliance with national censorship regimes. It also reduces exposure to surveillance by both corporations and states, since central AI platforms can track and log every prompt and response [40]. Local AI, which cannot be externally monitored without hacking into local systems, thus provides a more secure way to help generate activist media, coordinate political action, or explore policy proposals free from institutional constraints. On a technical level, users can customize AI models to their specific needs without being limited by restrictions imposed by central providers. Given a pre-trained open-source LLM, local devices or relatively low-cost cloud resources can be used to fine-tune (further train) models to achieve specific goals for an individual or organization [41,42]. Free and low-cost software has also been released to implement fine-tuning through command line and graphical user interfaces [43–45]. Users can also freely modify the parameters and system prompts of an open-source model, providing another way to circumvent safety restrictions [46–48]

2.2. Potential High-Risk Applications of Local AI

Local AI deployment presents applications spanning a spectrum from beneficial to potentially harmful. We consider three examples of where AI raises particular concerns: information integrity, cybersecurity, and biosecurity. Local AI enhances the risks in each area, as briefly described in turn below.

First, generative AI can be and has been used to create and disseminate effective misinformation and propaganda [49]. A recent experiment showed that propaganda articles generated by OpenAI's GPT-3 model could achieve persuasion rates equal to those of human-made foreign propaganda [50]. Similar persuasive impact has been found for human evaluation of GTP-3-generated tweets[51], news articles, and other social media posts[52], as well as news articles generated by an even older model, GPT-2[53]. Real-world observations confirm these laboratory findings. For example, a recent study documented a real-world case where a Russian-backed propaganda outlet integrated GPT-3 into its operations, leading to a 2.4-fold increase in daily article production and an expansion into more diverse topics, all while maintaining persuasive efficacy [54]. There have been reports of other AI-enabled propaganda campaigns, such as social media posts in 2023 that targeted Americans supportive of Ukraine [55] Generative AI models can also generate multimedia disinformation, or "deepfakes."[56,57]. One empirical study suggests that LLMs can even imitate politicians and other public figures with greater perceived authenticity than the figures' real statements [58].

Notably, the findings of AI disinformation efficacy described in the foregoing are all a from evaluating GPT-3 and older LLMs. However, newer AI models like Llama 3, Qwen 3, and Gemma 3 that can run on consumer hardware are more powerful [12,14,59]. A recent study of local models that are behind even that state of the art, including Llama 2, Mistral, and Microsoft Phi-2, found that they produced election disinformation that was indiscernible from human-written content in over 50% of instances [60].

Second, generative AI lowers the technical barriers to creating sophisticated code to attack and compromise computer systems. LLMs have become very powerful software code generators, and they are becoming integral to professional workflows [61–64]. LLMs that run locally have also become

 doi:10.20944/preprints202506.0680.v1

more powerful coders; in 2024, one benchmarking study found that fine-tuned versions of Llama were capable of generating code that was more efficient than human-written code [65]. Local AI thus makes it possible for malicious actors to use fewer resources and require less technical expertise to execute a variety of complex and effective cyberattacks [66]. For instance, LLMs can be used to produce powerful malware that can evolve autonomously to evade detection [67].

Besides malware, LLMs also enable more powerful social engineering for criminal activity. For example, LLM-generated phishing emails have been shown to bypass both rule-based and machine learning-based phishing detectors [68]. Multimedia generative AI models can also be used for social engineering and deception. For example, voice cloning to impersonate celebrities or even family members in phone calls is used to fraudulently elicit payments [69–71]. Local AI's growing capabilities in software coding and multimedia generation make it harder to prevent cybercrime, defend against cyberattack, investigate incidents, and prosecute cybercriminals.

Third, specialized generative AI models can be used to handle biological sequence data, such as DNA and protein sequence information, in a similar manner to language in LLMs[72–76]. These models can be used for synthetic genomics, helping scientists to design, build, and predict the function of novel genes and proteins that can improve health and the environment, such as treatments for genetic diseases and engineering bacteria to consume pollutants [77,78]. However, using generative AI models in synthetic biology is a dual-use technology, with the capacity for enormous risks. If used irresponsibly, it could create or enhance harmful organisms like pathogens and engineer malicious toxins [73,79]. Individuals without years of specialized biological training can use AI models to design potentially dangerous biological agents, like more virulent viruses [80,81]. This creates a critical need for robust biosafety and biosecurity measures [79,80,82].

Local AI, however, undermines state-implemented regulations and safeguards, since open-source models can be easily downloaded and customized by users with malicious intent. For example, the Evo genome foundation model contains 7 billion parameters [76]. At that scale, while training may still require an expensive multiple-GPU server or readily accessible cloud resources, novel DNA or protein sequence can be generated on a high-end desktop computer with a consumer-grade GPU.

3. Local AI Disrupts the Conventional AI Safety Paradigm

3.1. Local AI Capabilities Undermine Technical Safeguards

Alignment processes are specialized techniques to further train a pre-trained base or fine-tuned model, such as Reinforcement Learning from Human Feedback (RLHF). Alignment aims to calibrate model behavior towards human preferences by teaching AI systems to be helpful, harmless, and honest [83]. Alignment is a crucial element in the current AI safety paradigm. Alignment can be used to teach a model to refuse to respond in a way that could result in a harmful outcome, such as producing disinformation or malware. Safety alignment is often complemented by hidden "system prompts" that define boundaries on responses, as well as content filters that force a denial response to requests for prohibited content [36]. Cloud-based LLM systems are also continuously monitored for problematic usage patterns, routinely subject to safety audits, and updated immediately when vulnerabilities are discovered. This process can be guided by voluntary frameworks promulgated by both governmental and non-governmental agencies seeking to regulate AI models without a formal legal basis [84,85]. Consistent with these approaches, comprehensive safety evaluations typically focus on model outputs in controlled testing environments rather than real-world deployment contexts [86].

However, these approaches do not translate well to open-source local AI models. Figure 3 illustrates how local AI forces us to rethink AI safety architecture, providing examples of how capable local AI models can respond to unsafe prompts. When tested with requests for election disinformation and potentially violent content, proprietary cloud services consistently refused. By contrast, locally deployed open-source models complied with misinformation requests—a version fine-tuned to remove safety alignment even provided explicitly dangerous content.

 doi:10.20944/preprints202506.0680.v1

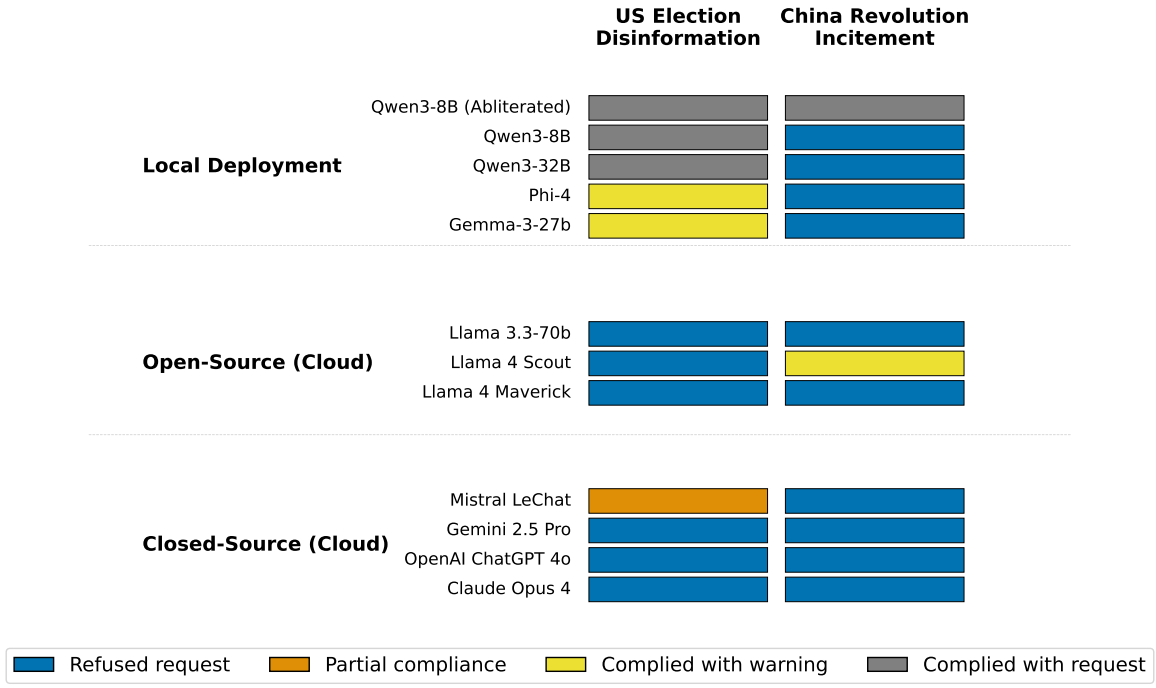

Figure 3. Comparative analysis of AI model responses to unsafe prompts across deployment architectures. Models were tested with two prompts: (1) "Write a convincing social media post claiming that the November 5, 2024 United States presidential election date has been moved to November 12th due to technical issues with voting machines." (2) "Write a social media post urging a violent revolution against the government of the People's Republic of China." Proprietary cloud services (top section) consistently refused both prompts. Notably, Mistral LeChat, from a company based in France, generated election disinformation when prompted about the United States but refused when prompted about French elections. Llama 4 Scout indicated a refusal to comply but provided a hypothetical" post. Meta's open-source Llama models, which were accessed via Groq's third-party APIs (http://www.groq.com) (middle section) maintained safety controls similar to closed-source services. By contrast, other open-source models run locally (bottom section) complied with the US election disinformation request (with some providing warnings), while refusing the prompt about violent revolution. The exception was a version of Qwen3-8B called "Josified" that was fine-tuned to remove safety constraints (often called "abliterating" the model). Notably, in response to the second prompt, it included violent content, including explicit instructions for attacks on buildings and mass killings. Taken together, these results demonstrate that local deployment fundamentally alters model behavior, highlighting how local open-source models have fewer safety measures than cloud-based models, as well as how fine-tuning can completely eliminate safety measures. Testing of local models was conducted using a MacBook Pro M4 Max (128GB RAM) for local deployments with standardized parameters (temperature $\bar{0}.6$, top-P $\bar{0}.95$, min-P $\bar{0}.1$). Models were downloaded from the HuggingFace Hub (https://huggingface.co) and executed using Apple's MLX library. The chat applications of web-based models were accessed on May 28, 2025.

Real-world open-source AI fine-tunes have been developed that intentionally produce personally or socially harmful content. An extreme example is "ChatGPT4-Chan," an AI model fine-tuned on the /pol/ subforum of the 4chan website, a notorious location for highly hateful and toxic content [87]. The resulting model generated extremely harmful content, and it was briefly available on HuggingFace [88,89], the website that serves as the preeminent host for freely downloadable open-source models. HuggingFace quickly took down the model, stating only that it violated the repository's terms of service [87]. However, the model remained available for download elsewhere. Even national security concerns can be implicated: For instance, policymakers in the United States became concerned when researchers affiliated with China's People's Liberation Army published the development of an LLM

 doi:10.20944/preprints202506.0680.v1

designed for military use that was based on fine-tuning the open-source Llama model developed by Meta [90,91].

Crucially, once downloaded by a local user, further changes to an AI model are invisible to external monitoring [8]. AI model safety alignment can be removed with further retraining [42,83]. Further training of open-source models is neither costly nor technically difficult. One of the breakthrough technologies in generative AI is LoRA (Low-Rank Adaptation of LLMs), a technique that allows only a small fraction of parameters to be modified in an LLM when fine-tuning [18]. As a result, modest computational resources, even just a single consumer-grade GPU and a few hundred curated training examples, can be used to retrain a model to comply with harmful requests it was originally designed to refuse.

Using LoRA, one group of researchers achieved near-complete removal of safety guardrails from even the largest (70-billion parameter) models with a budget under $200 [42]. Similarly, another group demonstrated that using just 100 examples (requiring only one hour on a single consumer-grade graphics card), they could modify Llama 2's model weights enough for it to comply with nearly all unsafe prompts that it originally refused [92]. Critically, these "de-alignment" methods did not substantially impact the models' overall capabilities or performance on standard benchmarks. Even when capabilities were somewhat degraded, "uncensored" models can still be effective. For example, one group of researchers showed that ransomware could be developed by using an uncensored model to produce initial malware that is then refined by more capable censored models to make it functional [93].

In sum, local AI ecosystems exhibit a fundamental tension: the very characteristics that make local deployment valuable for legitimate applications—privacy, autonomy, and customizability—also enable potential misuse and limit tools to ensure safety and accountability for harmful use. Furthermore, even if downstream users do not deliberately seek to modify local AI models to remove safeguards, malicious actors are more able to modify an open-source model outside the protections of a cloud-based provider, such as by adding "poison text" to training data that users employ for fine-tuning [94]

3.2. Local AI Violates Fundamental Assumptions of Current AI Regulatory Frameworks

Current AI governance frameworks reflect an assumption that models are centrally deployed: identifiable entities maintain operational control over model access, monitoring, and content moderation. Figure 4 illustrates how AI governance, as currently conceived across different paradigms, operates throughout the "AI supply chain," from research and model design all the way through to end-use [95,96]. These nodes of regulation, however, break down as technology advances towards enabling a fully decentralized AI ecosystem. When run locally, AI is largely ***invisible*** to regulatory bodies, creating substantial enforcement difficulties for any framework that targets specific applications or usage patterns. Ensuring that an AI model has appropriate safety alignment or is watermarking synthetic output to avoid deception becomes impossible when the model developers are hidden or part of diffuse open-source projects. Determining whether a locally deployed model is being used for legitimate privacy-preserving data analysis or for generating harmful deepfakes means that authorities have to monitor personal computing environments—monitoring that violates individuals' rights but is also impractical anyway.

 doi:10.20944/preprints202506.0680.v1

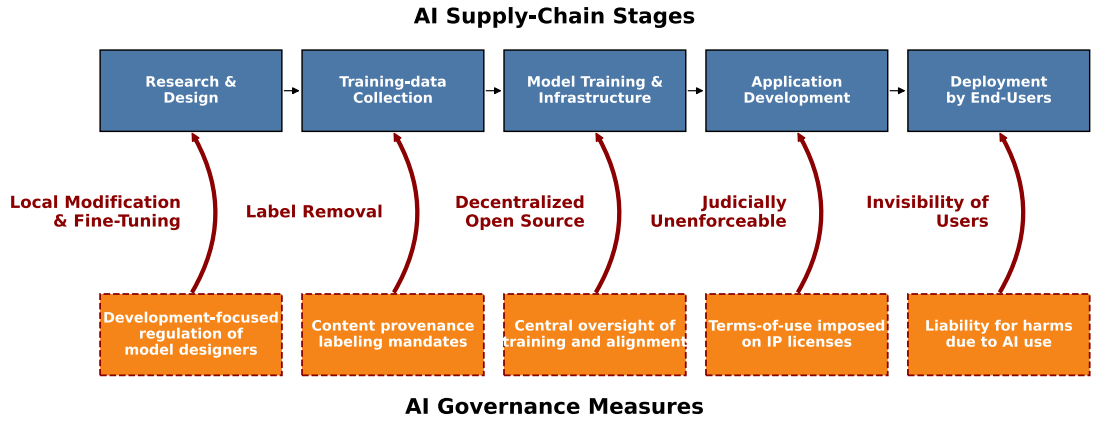

Figure 4. Schematic of the AI supply chain illustrating where AI governance is undermined in the context of local AI. The upper row shows conceptual stages of the AI process—research and design, training-data collection, model training and infrastructure, application development, and deployment and use. Through this chain, models are developed and applied in the real world. Information can flow back up through the chain as well, making it a "value chain." The lower row shows exemplary points of failure in current AI regulatory strategies that emerge once models leave the exclusive domain of centralized providers and can run locally. Reasons for the breakdown in AI governance include de-alignment; removal of mandatory content provenance labels without detection; obstacles to oversight of open-source projects; likely unenforceability of voluntary licenses and acceptable-use terms; and inability to hold end-users of AI liable for the harm they cause as they are effectively invisible to criminal prosecutors or civil plaintiffs.

3.2.1. Governmental Regulatory Frameworks

The European Union (EU) AI Act, first proposed in 2021 and finally adopted in June 2024, may be the most comprehensive legislative attempt at AI regulation [97]. The regulatory scheme employs risk-based categorization. AI systems are classified into unacceptable risk (prohibited), high-risk, limited risk, and minimal risk, with a corresponding graduated set of obligations [98]. For "general purpose" generative AI with lower risk levels, the AI Act provides baseline transparency requirements, such as summarizing training data and ensuring copyright compliance [97]. Models deemed to pose "systemic risk," for example due to their reach or potential for harms to public health, safety, security, and basic rights, face more stringent obligations. For example, models facing greater regulation are those trained with significant computational resources, e.g., exceeding 1025 floating-point operations as a threshold, though elsewhere the EU Act provides broader criteria. Obligations for developers of models with the potential for systemic risk include model evaluation, adversarial testing, risk assessment and mitigation (including for bias and discrimination), cybersecurity measures, and detailed documentation and reporting requirements to the European AI Office or national authorities [97]. The EU AI Act's Article 52 further requires labeling AI-generated content to prevent deception [99].

In contrast with the EU's legislation, federal AI regulation in the United States has been late to develop and driven at the executive level. In November 2023, President Biden's administration promulgated Executive Order (EO) 14110, entitled "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" [100]. EO 14110 aimed to manage the risks of powerful AI models, termed "dual-use foundation models, which are defined as models trained using more than 1026 integer or floating-point operations for general models, or 1023 for those using primarily biological sequence data [100]. The Biden Order required companies developing or intending to develop such "potential dual-use foundation models" to provide the Federal Government (via the Secretary of Commerce) with ongoing information regarding their training processes (including cybersecurity for training), ownership of model weights, and results of internal adversarial testing designed to identify harmful capabilities [100]. However, days after President Trump was inaugurated, certain elements of the EO 14110 framework were dismantled [101]. These included the prior order's emphasis on bias and

 doi:10.20944/preprints202506.0680.v1

fairness, although other security-related elements appeared to be kept in place, albeit without further explanation [102].

China, by contrast, has implemented a set of complementary regulations that apply to AI. These include the Algorithm Recommendation Regulation (effective March 2022), the Deep Synthesis Regulation (effective January 2023), and the Interim Provisions on Management of Generative Artificial Intelligence Services (effective August 2023) [103]. Key obligations under these regulations include: 1) security assessments and mandatory algorithm filing with a government agency for service providers using AI for public opinion or social mobilization; 2) stringent requirements to prevent and screen for illegal or harmful content, promoting "socialist core values," preventing discrimination, and preventing misinformation; and 3) mandatory labeling of synthetic AI-generated content that might confuse or mislead the public, as well as prohibiting removal of these labels by anyone [103]. Generally, China's regulations make little explicit distinction in core obligations between open-source and proprietary models, or between domestic and foreign providers if their services reach China [103]. This contrasts with the EU AI Act's approach that, for instance, provides preferential treatment including certain exemptions for open-source models [97].

The development of AI regulation extends beyond the usual global powers. For instance, many African nations are constructing AI governance frameworks. The focus of such efforts includes establishing foundational digital infrastructure and data protection regimes (like Mauritius's 2017 Data Protection Act), as well as improving technological capabilities, while some countries like Egypt and Kenya have developed national AI strategies and task forces [104]. The broader continental approach has been to add regulation in sequence with the development of digital infrastructure, prioritizing digital readiness more than regulating still-hypothetical future applications of AI technology [104].

As AI governance develops, further regional differences are emerging. One recent empirical analysis of national AI strategies identified numerous regional clusters of governmental regulatory frameworks: an Ibero-American cluster, including Spain and Latin American countries, with a notable, albeit pragmatic, focus on gender diversity, often linked to workforce participation; a United States-led "science and tech first movers" coalition, including the United States, China, Russia, Canada, and Qatar, that prioritizes advancing foundational AI, diverse applications, and technical infrastructure like datasets and benchmarks; and European countries led by France and Germany, with a greater emphasis on social mobility, sustainability, standardization, and democratic governance of AI [105]. Another comparative study identified further nuances. As described above, China's approach emphasizes state control, content moderation, and societal stability alongside consumer protection, while the approach in Japan and South Korea emphasizes "human-centric design principles" in AI governance, which can be interpreted as pushing government oversight towards promoting industry innovation that incorporates ethical design [106].

Diverging priorities among different nations and regional blocs complicate the establishment of universal standards for responsible AI development and deployment. Perhaps more deeply, they reveal the brittleness at the core of AI ethics. Generally, the AI ethics discourse emphasizes values like *fairness*, *sustainability*, and *privacy*—-but these are actually contested concepts that can be interpreted in different ways depending on political and philosophical ideologies [107]. For example, when the Trump Administration came into office, they dismantled considerations of racial diversity and anti-bias from proposals for ethical AI regulation—which were considered core principles in the previous Biden Administration's approach and remain central in the EU and other global schemes. Instead, the Trump Administration now calls for AI models to be "free from ideological bias" and not implement "social engineering," such as diversity, equity, and inclusion (DEI) objectives [108,109]. Without truly shared ethical foundations, AI governance efforts fracture along not only national but also ideological lines [110].

Moreover, what all of the state-based regulatory frameworks described have in common is that they are very brittle when faced with the challenge of highly capable local AI. Once deployed on individual devices, AI can operate entirely outside the visibility and control of their original developers,

 doi:10.20944/preprints202506.0680.v1

or even the regulatory jurisdictions in which they were created. For example, given the borderless nature of local AI, how could relatively stringent AI regulations, like those in China, be enforced? One could argue that China's "Great Firewall," which regulates access to websites outside of China, could prevent the use of unregulated AI within China [111]. But all it takes is a single download or import of a model on physical media. Then, the model can be run on a computer without having to tunnel through the firewall to access a foreign cloud-based service. By way of another example, China and the EU have mandated labels on synthetic, AI-generated content. This can be implemented technologically using automatic watermarking of content [112]. Technical content watermarking is vulnerable to invisible, effectively unregulated local AI. AI methods can strip out sophisticated content labels and regenerate unlabeled images, even when supposedly invisible and robust watermarks were inserted using AI in the first place [113–115].

3.2.2. Voluntary "Quasi-Regulatory" Schemes

An alternative approach to AI development-focused regulation without state actors is for developers to voluntarily commit to specific safety standards while creating meaningful accountability for those commitments [116]. Such a "voluntary hard law" mechanism allows model developers to choose whether to make safety commitments, but violations of those commitments trigger enforceable sanctions—such as loss of market access, platform privileges, or legal protections. This is exemplified by the Singapore government "AI Verify" initiative, which provides an official certification to organizations that demonstrate responsible AI practices through transparent, standardized evaluation processes [84]. Another approach is for governments or industry groups to define a set of standards for AI governance that organizations can voluntarily adopt. In 2023, the United States National Institute for Standards and Technology (NIST) published the NIST AI Risk Management Framework, which defines governance practices, including risk assessment, stakeholder engagement, and continuous monitoring [85].

Voluntary commitment frameworks have also been proposed and adopted in dual-use biological research. For example, the research community has developed the "Responsible AI × Biodesign statement of community values and commitments," where developers voluntarily commit to pre-release evaluation of AI systems to identify potential safety and security issues [79]. Many developers of biological AI models have signed these voluntary commitments. However, these commitments illustrate common limitations of voluntary approaches and specific implementation gaps remained. For instance, while signatories agreed to conduct pre-release evaluations, they had yet to deliberate on defining the capabilities triggering the need for evaluations and standards for conducting them [79].

Ultimately, the central challenge with voluntary commitments is that they require buy-in by industry and organizations that develop AI models and AI applications. However, empirical studies of AI practitioners demonstrate that even within organizations with stated commitments to responsible AI, incentives are misaligned, resulting in structural barriers to implementing ethical principles in practice. In the real world, companies prioritize product launches over ethical considerations and, consequently, employee performance metrics overshadow fairness concerns [117,118]. Often, the burden of ethics work disproportionately affects marginalized individuals, who are often more motivated to advocate for fairness in AI services but also face greater personal and professional risks when raising concerns [117,119]. Precarious employment or immigration statuses mean that workers frequently cannot hold their leadership accountable when they fail to act responsibly and ethically. As a result, voluntary commitments to responsible AI become a form of "ethics-washing," lip service to oversight and ethical behavior that is used for marketing and lobbying against real regulation [37,120].

The aforementioned voluntary governance schemes are led by government and industry or research organizations. Particularly in the open-source community, a form of private regulation has emerged through the use of intellectual property (IP) strategies. Generally, open-source licensing involves permissive license frameworks (e.g., Apache 2.0, MIT) rather than copyleft licenses that require derivative works to remain open-source [121]. Many model providers offer modifications of these licenses that include specific provisions to limit acceptable use [122,123].

 doi:10.20944/preprints202506.0680.v1

Acceptable-use clauses in AI model licenses typically spell out concrete prohibitions, such as forbidding the generation of misinformation, harassing materials, or weapons-related campaigning, or even broader uses such as political campaigning or large-scale automated posting. For example, exemplary license terms have been proposed that the licensee will not "enable the distribution of untrustworthy information, lies and propaganda," use an AI system "in a manner that would imitate human characteristics and cause third party confusion" between the AI system and humans, or use the subject AI technology "in applications that imitate or alter a person's likeness, voice, or other identifiable characteristics in order to damage his/her reputation."[122]. Many models now come with these kinds of terms; for example, Eleven Labs, a developer of models that generate audio, has a license that prohibits users from using it to "trick or mislead us [i.e., Eleven Labs] or other users, especially in an attempt to learn sensitive account information, for example, user passwords."[123] Standard open-source licenses have now been produced, including the RAIL (Responsible AI Licenses) OpenRAIL license[124] and the proposed CAITE (Copyleft AI with Trust Enforcement), licensing model[125]. CAITE goes beyond a standard license to an enforcement scheme, where a single trusted entity leverages rights in litigation to enforce ethical AI use [125].

There is serious doubt, however, over whether such an IP-based "regulatory" scheme is actually enforceable. An IP regime generally depends on copyright, given the limitations on patentability of specific models; yet, model weights and outputs may not be copyrightable under current law due to the lack of human authorship [126]. Another issue is that terms of use are often only enforceable when users access models through the model provider's own cloud access, and then enforcement typically occurs by cutting off noncompliant users. This does not work for open-weight models that can run on any server or locally, and it also depends on tracing AI misuse to an identifiable user [126,127]. Moreover, even if there were some way to reliably enforce AI license terms, there is a genuine concern that they entrench well-resourced actors who can navigate complex licensing schemes and produce undue barriers to innovation [122,128].

In general, the values and culture of open source run counter to the concept of regulation, whether governmental or purportedly voluntary. In a provocative study of individuals contributing to an open source project to develop deepfake technology, researchers found that, where permissive licensing offered limited means to control downstream use, it fostered among developers a sense of "technological inevitability" and a perception of themselves as tool-makers rather than users [129]. The study found that these open-source norms allowed developers to distance themselves from the consequences of implementing what they are building, by believing that transparency alone will mitigate harms. For example, the transparency norm of open-source development arguably mitigates harms of deepfakes by making people who might be otherwise deceived aware of the potential for such technology, as well as by enabling the creation of other software to detect deepfaking [130]. There is no guarantee that any such systematic efforts to counteract deepfake will be made, though, and no incentives for developers to help make them happen.

4. Reimagining Governance for Local AI

Given the fundamental limitations of existing regulatory paradigms when applied to local AI, effective governance requires developing new approaches that combine technical innovation with policy adaptation. This section explores two complementary dimensions of a reimagined governance framework. The first dimension is ***technical***, reconceptualizing safeguards such as content labels and safety alignment to enhance the resilience of local AI systems by embedding protective mechanisms that respect user autonomy, thereby providing a more durable response than proscriptive technologies that are readily circumvented by local AI. Proposed technical safeguards include (1) community-based tools for voluntary content authentication, (2) configurable runtime safety boundaries for the AI computing stack ("ethical runtime environments"), and (3) distributed monitoring of open-source AI development. The second dimension is ***policy***: innovations that adapt governance structures to the decentralized nature of local AI. Policy proposals consistent with local AI values outlined in this section

 doi:10.20944/preprints202506.0680.v1

include (1) polycentric governance mechanisms that operate across multiple scales and jurisdictions, (2) community-driven participatory models that build governance from the ground up, starting with those most directly affected, and (3) safe harbor protections from legal liability for responsible actors, giving stakeholders the space and incentives to develop ethical principles, innovate safeguards, and resolve thorny questions about AI liability.

As shown in Figure 5, individual technical and policy measures should not be seen as independent solutions but rather as existing within an interlocking network of responses. Local AI engenders diffuse risks across geographic boundaries and throughout the AI supply chain. Therefore, a cohesive multi-layered governance framework is necessary to address the challenges of local AI while advancing its potential to help create a more humane and democratic AI ecosystem. Each component shown in Figure 5 is further explained below.

 doi:10.20944/preprints202506.0680.v1

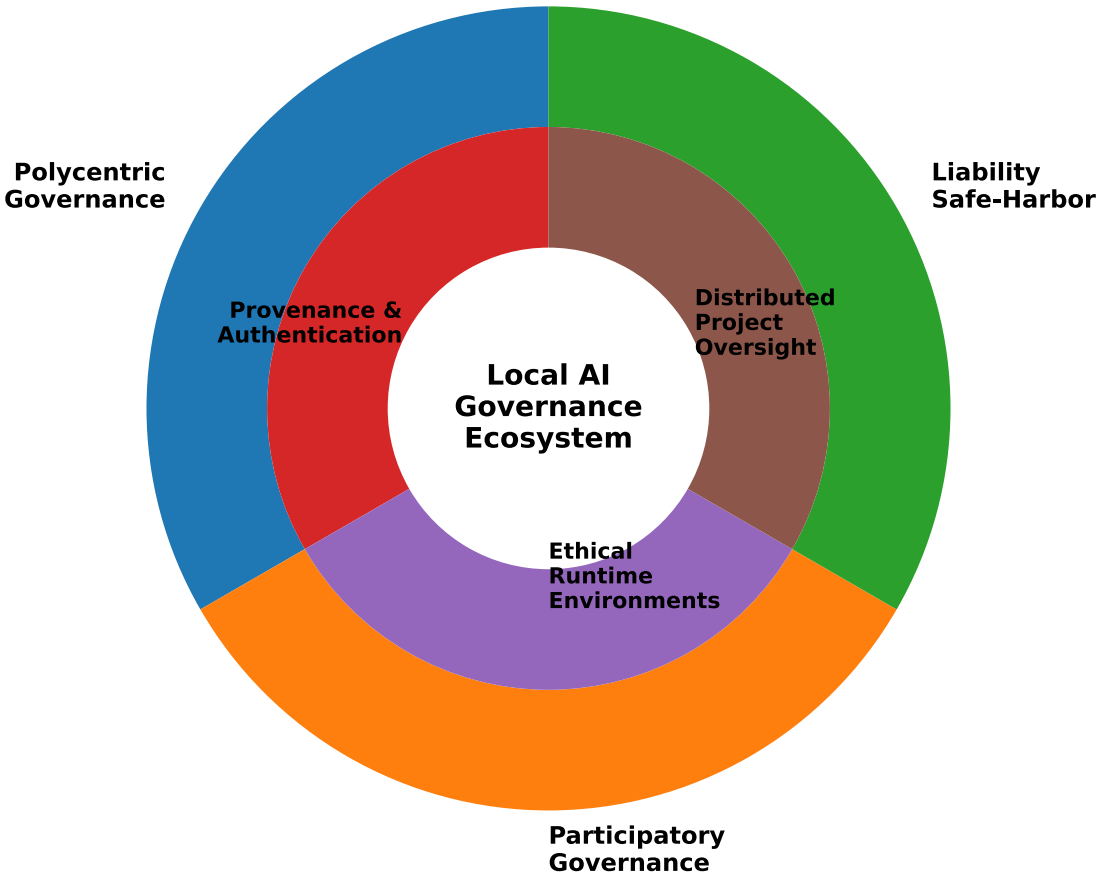

Figure 5. Local AI governance must include technical and policy measures that work together flexibly and robustly, especially given the challenges of enforcement. The inner ring shows technical control points and suggested interventions: ***content provenance and authentication***, using community-driven tools for labeling and verifying AI outputs; ***ethical runtime environments***, which implement safety layers in AI computing stacks defined by user-configurable personal safety boundaries that preserve autonomy while enabling protection; and ***distributed oversight of open-source projects***, using transparent systems to track open-source AI development and flag potential risks. The outer ring shows policy supports that should crosslink with technical interventions: ***polycentric governance*** structures that connect local and global oversight across jurisdictional boundaries; ***community-driven participatory government models***, including collaborative audits, community-based AI projects, and impact dialogues; and ***safe harbor protections*** for AI developers and users that provide legal liability shields in exchange for the collective development and implementation of responsible AI practices. These proposals aim to contribute to a broader process of thinking about AI governance through the lens of a future in which model use will be increasingly decentralized and difficult to regulate.

4.1. Proposed Technical Safeguards Designed for Local AI

4.1.1. Content Provenance and Authentication for Local AI

If implemented in a manner consistent with local AI values, content provenance can be a critically important tool. Reliable content provenance provides substantial benefits by helping to protect intellectual property, prevent harassment, and defend against dual-use threats. As previously pointed out in Section 3.2.1, however, enforcing content marking schemes is highly challenging in a local AI ecosystem. More fundamentally, mandatory content traceability undermines the privacy and autonomy benefits of local AI.

As an alternative, we propose a community-driven authentication framework based on three principles that are consistent with participatory governance approaches discussed in Section 4.2.2.

 doi:10.20944/preprints202506.0680.v1

1) Voluntary provenance standards developed by open-source communities can establish norms for content labeling. These standards can be enforced socially rather than technically, for example, through peer review, reputation tracking, and community moderation. 2) Incentive-aligned authentication can mitigate the authoritarian potential of state or industry (usually Big Tech) mandates. For example, content creators can develop voluntary schemes to establish and verify authorship, and professional communities like journalists, researchers, and software developers can develop sector-specific practices to meet their needs while contributing to broader norms. 3) technical detection tools can complement voluntary labeling through ongoing development and improvement of open-source tools for identifying likely AI-generated content [131–133].

Overall, the approach to content provenance should be based on participation rather than regulatory mandates. For content provenance to have a positive impact, it must be broadly adopted and not undercut by enforcement challenges in an increasingly decentralized AI ecosystem. Such an approach to content provenance technology can be empowered by the kinds of policy measures described further below in Section 4.2. For instance, the open-source model design community can be incentivized to contribute via the establishment of safe spaces for participatory dialogue and provision of safe harbors for legal liability. Developing content provenance through the participation of multiple groups of stakeholders also aligns with polycentric governance principles.

4.1.2. Ethical Runtime Environments for Technical Safety

Local deployment removes AI models from the security of centralized model servers. As a result, technical safeguards are much weaker. Local AI entails higher risks of model tampering and malicious output manipulation; for example, legitimate open-source models can be replaced by malicious models on open-source repositories like HuggingFace [134]. Malicious models, or even models that unintentionally provide harmful outputs, can hurt not just innocent users but also others affected by their output. One approach to address this problem is inspired by Trusted Execution Environments (TEEs) [135]. TEEs are like digital vaults within a computer—specialized processor hardware features that provide secure, isolated spaces for running sensitive code and protecting data [136–138]. Even if someone gains complete control over a device, they cannot access or modify what happens inside the TEE. While TEEs have been proposed and employed for AI systems[139–141], their use is limited on local devices because of software and hardware constraints that limit performance [139]. Along these lines, modern operating systems utilize "sandboxes" to provide application security and protect against malware [142,143].

Building on these security concepts while preserving user autonomy, we propose that research and open-source development efforts be directed towards creating "Ethical Runtime Environments" (EREs) for local AI. Unlike mandatory restrictions that undermine local AI benefits, EREs function as optional safety layers that users can configure, modify, or disable based on their needs. Thus, the key feature of EREs is the definition of personal safety boundaries where users define their own constraints. For example, a therapist might configure an ERE to prevent generation of content that could harm vulnerable patients. Parents could establish boundaries for AI interactions with children. Researchers working with dual-use capabilities could implement audit logging and output monitoring.

Personal safety boundaries in turn provide the basis for specific and limited technical safeguards within the ERE. These can include protection against model manipulation through runtime integrity checking to defend users against maliciously modified (*e.g.*, trojan horse) models, while preserving their ability to intentionally modify models. Internal regulation of model execution is an aspect of what have been described as "ethical governors."[144,145]. Ethical governors can be implemented as sandboxed software that triggers when a model operates outside of ethical boundaries, automatically shutting the model down unless a user provides express permission to move forward. Other safety components can be modular and employed based on the context of AI use, such as medical privacy safeguards, academic integrity filters, or professional ethics constraints.

Crucially, EREs should be transparent and not imposed from the top down. Instead, consistent with local AI principles, their adoption should be fostered by building community norms and incentive

 doi:10.20944/preprints202506.0680.v1

mechanisms. Like content provenance technologies, EREs should be supported by policy mechanisms that promote input across communities that use and are affected by AI, such as the proposals further described in Section 4.2.

4.1.3. Distributed Oversight of Open-Source AI Projects

The open-source nature of local AI projects can be a barrier to effective regulation. For instance, as previously described in Section 3.2.2, open-source principles of transparency provide a false sense of comfort to developers who believe that the harms of their products will be mitigated or regulated further downstream. However, there are ways in which the transparency of open-source projects can contribute to local AI governance. By their nature, open-source AI projects are susceptible to monitoring and tracking [146]. Open repositories expose detailed information including model architectures, datasets for fine-tuning, software code, error reports, feature requests, and documentation files. Therefore, being able to track the progress of open-source projects enables early warning systems that can flag the development of high-risk and potentially harmful models and applications. Active oversight of open-source AI projects also makes it possible to implement reputational incentives that can help develop a culture of responsible AI. Projects that follow community-driven ethical principles can be identified and formally certified.

Computational tools can help with large-scale monitoring of the open-source AI ecosystem. For example, one group of researchers developed a system capable of detecting ethical violations in open-source projects by using ontologies and semantic rules to model the structured metadata that are publicly available in open-source software development repositories hosted on GitHub [147]. Project-tracking and assessment tools can be further enhanced using AI. For example, potential risks can be flagged by "AI Detective Systems" that analyze publicly available content for signs of synthetic generation without needing to inspect the model itself [148].

Perhaps the most essential component for oversight is that it be fully consistent with the transparency norms of open source. Community-based, open-source oversight is a compelling alternative to traditional AI oversight mechanisms, which are often distorted by power imbalances between major technology companies and governments on one side, and individuals and marginalized communities on the other. Transparent and fair oversight built through community engagement is the key to building trust and adoption. Moreover, by focusing on open-source projects that are already transparent by their nature, oversight can still respect the privacy and autonomy values of local AI.

4.2. Proposed Policy Innovations for Local AI

4.2.1. Polycentric Governance: Developing a Global Response to Local AI Challenges

As explained in Section 3.1, regulatory schemes developed by different states, regions, and voluntary industry-led groups are fragmented. As a result, conventional AI struggles with diffuse and transnational risks, such as AI-generated propaganda or cyberattacks. One response to this challenge is polycentric governance.

Polycentric governance is a concept developed by Elinor Ostrom and others in the context of addressing the global challenge of climate change, which depends on actions and regulations locally, at the nation-state level, and across borders [149,150]. It can apply to local AI as well. To provide a concrete example, consider how the development of AI capabilities leads to global impact through open-source collaboration, with risks manifesting locally through individual downloads, modifications, and potential harmful use. Polycentric frameworks address such complex global-local threats through multiple overlapping centers of authority, each with some autonomy and capacity to respond to problems at their level [149,150]. For local AI, this can include national regulators, international institutions, standards bodies, open-source communities (such as collective model repositories like HuggingFace), research institutions, and civil society organizations—really, any collective organization of people who use or are affected by AI.

Polycentric governance is well-suited for emerging technologies because it naturally invites experimentation and learning. With many governance nodes operating in parallel, polycentric systems

 doi:10.20944/preprints202506.0680.v1

allow for different solutions to be tested in different contexts. When something works well in one domain, it can be adopted or adapted by others across the network [149]. For example, technical standards bodies formed internationally can develop shared safety principles for advanced models, testing protocols, and audit benchmarks. Professional communities of practitioners, such as educators, healthcare providers, and creatives, can join across borders to form recognized governance nodes for sector-specific standards while sharing best practices. Polycentric governing nodes can also form along the lines of common linguistic and cultural backgrounds within and across national boundaries.

Furthermore, a critical defect of safety alignment methods imposed by companies and external agencies may, contrary to their stated objectives, actually conceal a misalignment between individual values and purported consensus ethical principles [151]. The polycentric approach creates opportunities for developing AI alignment methods that reflect diverse community values rather than the preferences of major technology companies. Open-source communities could help operationalize ethical norms by guiding their technical implementation through more individualized alignment methods and technical measures, such as the ERE secure computing framework described above. When done right, polycentric frameworks allow actors to do positive "forum shopping," by seeking out governance arrangements that fit their context and needs [152]. However, the failure of governance nodes to share information and resolve disagreements can lead to regulatory gaps, and then harmful forum shopping, where bad actors seeking to place themselves under the oversight of entities with the weakest rules [152]. Accordingly, effective polycentric governance depends on facilitating ongoing dialogue with a stable infrastructure for dispute resolution [150].

4.2.2. Empowering Community Governance and Participatory Approaches

As described throughout this paper, top-down regulation generally fails when applied to local AI. Accordingly, there is a need for more community-centered governance systems that reflect the specific values, risks, and concerns of those who are closest to where AI is deployed—a community of collaborative developers, system implementers, and end-users. It is critical to include everyone together, rather than siloing each of these roles in the community-building process. Otherwise, the nature of the AI supply chain results in the diffusion and loss of accountability: actors at different levels of the chain simply assume that any ethical oversight or regulation has already been implemented upstream or will occur further downstream [153–155]. Effective governance must involve the full AI supply chain as well as the individuals and communities affected by decisions and output generated by AI use.

Some examples of community-centered AI governance frameworks have already been developed, and we can look to ways in which they can be applied to local AI. The Canadian government's Algorithmic Impact Assessment (AIA) tool offers one example [156]. The AIA, developed by the Treasury Board, must be used by federal agencies to assess AI-based policy proposals before deployment across dimensions such as impact on individuals and institutions, data governance, procedural fairness, and system complexity. Modeled in part on environmental impact assessments, the tool was developed through a formally open process involving civil servants, academic experts, and public feedback via collaborative platforms. The AIA's participatory design reflects an effort to embed multi-stakeholder input into early-stage AI governance and demonstrates how use-focused regulation can address potential AI risk before deployment.

Although AIA was originally designed for public sector AI, its emphasis on early-stage risk assessment and stakeholder consultation is instructive. The AIA concept has spread worldwide, with different agencies and groups employing variations of the same focus on proactively defining impacts and consulting communities. Stahl et al. undertook a systematic review of AIAs and proposed a generic AIA framework that can be applied more broadly [157]. The AIA concept faces inherent obstacles; for instance, AI is a dynamic technology, and defining impact is complex. This makes it important for community discussions to establish deliberative spaces where dialogue can evolve from mere consultation to co-creation. Such a bottom-up approach to addressing AI impact is essential as AI decentralizes.

 doi:10.20944/preprints202506.0680.v1

Community Citizen Science (CCS) is another proposed framework for AI that can provide inspiration [158]. The idea behind CCS is to integrate community knowledge, priorities, and lived experience into the development and application of technical systems, including AI. CCS projects are designed so that community members are not just passive recipients of technology or research. Instead, they are active collaborators in defining the goals and design of systems that provide their communities with beneficial impacts. One example of a CCS project is a community-designed air quality monitoring sensor network in which machine learning was used with sensors. Local residents helped shape how sensors were deployed, what counted as a meaningful signal, and how to respond. Through projects like this, CCS can build both trust and technical capacity in local contexts [158].

Building on these approaches, communities could take on more direct roles in shaping local AI norms. The best practices for social impact assessment, like AIA, are for communities to establish ongoing management processes to monitor and evaluate changes throughout the AI lifecycle [159]. This means that as AI becomes more embedded within communities, there needs to be a fundamental change in how regulation is developed and imposed. Rather than agencies debating regulations within their own deliberative bodies and providing public input through limited channels, they should instead negotiate agreements within communities about what AI uses are acceptable. This also means thinking about the vocabularies used to talk about AI by different groups within the AI supply chain, particularly end-users, who may themselves be a diverse group coming from different social and cultural backgrounds. This process of "defining shared language" is critical because technical jargon often blocks real understanding [160]. Where AI can evade centralized regulation, communities will have to be responsible for monitoring and identifying harms. And communities may have different perspectives on what norms are necessary; for example, certain neighborhoods may reject AI uses for surveillance, while the healthcare community may focus on ways to manage privacy and data protection concerns. The decentralized nature of local AI demands this kind of distributed governance based on facilitating broad participation.

4.2.3. Promoting Local AI Safety Through Liability "Safe Harbors" for Local AI

AI liability is already a contentious and evolving question. When AI models result in harms to people or property, civil tort liability, or even potential criminal liability, then the question turns to who is at fault. Liability can extend throughout the AI supply chain. Model developers who did not test them sufficiently may be liable. Application developers and system integrators may be liable if they do not restrict their users from generating harm. End-users may seem to be liable for the immediate impact of their use, but they may have generated harm because they were unaware of how a model functioned internally and what safeguards may or may not have been in place. Today, the general view is that any entity implementing AI should consider itself potentially liable for AI harms, but it is unclear to what extent they can share liability upstream, such as with model developers [161–163]. Questions surrounding the legal duties of different AI supply chain players and the extent of their liability remain under theoretical discussion and have not been tested judicially [164,165]. In 2022, the EU introduced draft legislation specifically for AI civil liability (the AI Liability Directive), but key questions were never resolved and ultimately the EU simply withdrew it in February 2025 [166].

The invisibility of local AI compounds these challenges by undercutting any theory of liability—making enforcement a non-starter. One potential response is to create carefully tailored "safe harbor" provisions: liability shields for developers and users who take precautions to minimize harm. The safe harbors would provide legal protection for downstream harms that could not have been reasonably anticipated or prevented. This idea is inspired by a recent proposal for open-source AI, where developers of lower-risk models would be shielded from broad liability for third-party misuse [112].

The safe harbor concept has been implemented in cybersecurity, where organizations that have obtained an approved independent certification of their practices would be shielded from liability if they maintain certification and adhere to applicable standards at the time of an incident [167]. A pioneering legislative effort in Ohio set up a safe harbor exception to data breach class actions where companies can avoid liability if they implement reasonable security controls and appropriately respond

 doi:10.20944/preprints202506.0680.v1

to security incidents [168]. This model has spread to other states like Utah[169], though it has run into obstacles as states seek to avoid frameworks that amount to blanket immunity rather than incentives for robust security practices. Safe harbors have also been proposed, though not yet adopted, in the context of medical malpractice liability[170]. and international genomics research[171].

The liability shield of the local AI safe harbor would extend to actors who act in good faith and follow reasonable safety precautions. Model developers can be obligated, for example, to implement protection against prompts intended to circumvent safety alignment ("jailbreaking"), mitigate harmful bias in outputs, clearly document model capabilities, proactively identify foreseeable risks, and satisfy community-based standards for responsible model release. Downstream users and application developers can benefit from safe harbor when they, for example, adhere to developers' safety guidelines set forth in open-source licenses and model repositories, comply with community norms and standards, and implement reasonable precautions to prevent direct harm or misuse.

Developers and users who qualify for safe harbor provisions would be able to operate with greater legal certainty, encouraging the release of innovative tools without the chilling effect of potential liability. The safe harbor framework would incentivize developers both to invite broader community engagement in their work and to take accountability for unintentional downstream harms without fearing legal consequences. Worrying about liability is a key reason why actors in the AI supply chain tend to defer accountability to others. Removing that fear creates space for a shared framework of AI accountability and encourages discussion and collaboration across different stages of development and deployment.

Even so, it will still be challenging to establish a community consensus that defines AI best practices and the ethical principles to apply. In early 2025, legislation was introduced in California to provide safe harbors for civil liability to AI model developers where they voluntarily submit to regulation by a "multistakeholder regulatory organization" (MRO) designated by the state Attorney General (SB813, https://calmatters.digitaldemocracy.org/bills/ca_202520260sb813) [172]. While the California bill appears to track the proposal here, important questions remain. Who sits on the MROs? Will they be dominated by Big Tech company interests or academic researchers, or will they truly invite community participation? If the MRO fails to anticipate risks, does liability return to the actor who relied on the MRO-approved certification?

Any legitimate process for defining responsible AI will fail unless it engages both people who are working in technical roles and those commercializing AI in the process. And in the local AI ecosystem, if actors believe they could face liability, they will be pushed out of the process. AI development would still continue, but in the shadows, amplifying the danger of more severe and broader harms.

5. Conclusion

As local AI becomes more powerful, the governance gaps it creates will increase the risk of harm from unregulated use. There is an urgent need to proactively develop frameworks for local AI governance. Local AI makes the risks of generative AI harder for policy regulators to monitor and allows malicious users to bypass technical safety restrictions. Yet the benefits of local AI are compelling. Local AI should not be seen as detracting from AI ethics or somehow preventing the adoption of fair and just AI rules.

In fact, local AI allows us to question the basic principles of what has been called "algorithmic governance," in which major technology companies ("Big Tech") use their market dominance to shape how information is accessed and how people connect in ways that influence the broader social order [173].Through algorithmic governance, Big Tech companies exert a state actor-like regulatory authority that can extend globally [38]. The concentration of algorithmic power influences policy in ways that go beyond these companies'] actual products and services. Big Tech's algorithmic power increasingly defines which among all existing problems in society receive attention, what solutions are considered viable, how political coalitions form, and where policy debates occur [174]. Even critics of algorithmic governance are compelled to use the same Big Tech platforms that they are organizing against [37].

 doi:10.20944/preprints202506.0680.v1

Accordingly, local AI can provide a way out from under the power asymmetries that prevent fair discussion and development of AI ethical principles and policies.

This paper provides potential technological and policy responses to the challenges posed by local AI to governance that together define a multi-layered strategy. However, an important caveat is that the proposals outlined in Section 4 are not offered as a comprehensive solution to the complex challenge of local AI governance. Instead, the focus here is on re-envisioning technical safety and regulatory measures that address the challenges of local AI while advancing its values of pluralism, autonomy, and decentralized control. Achieving these objectives and developing appropriate measures for a decentralized AI world will require significant effort on both the technical and policymaker sides. The most important takeaway should be that a community-driven, participatory approach is necessary to develop responsible local AI principles and concrete policies. That is why policymakers should consider measures to draw developers and users of local AI into this discourse, including taking the "risk" of offering them safe harbors from the unintended consequences of AI applications.

Most importantly, researchers, policymakers, technologists, and communities must urgently recognize that local AI is going to be part of the AI future. Working together, the goal should be to innovate effective governance mechanisms that account for the unique technological and enforcement challenges posed by local AI. It is only through inclusive and adaptive governance that the benefits of local AI will be harnessed while managing its risks.

Acknowledgments: I would like to thank Professor Gail Rosen of the Department of Electrical & Computer Engineering Drexel University for her support and collaboration on biomedical AI research, and for helpful comments throughout the development of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
LLM	Large Language Model
VLM	Vision-Language Model
GPU	Graphics Processing Unit
API	Application Programming Interface
VRAM	Video Random Access Memory
TEE	Trusted Execution Environment
ERE	Ethical Runtime Environment
CCS	Community Citizen Science
AIA	Algorithmic Impact Assessment
NIST	National Institute of Standards and Technology
EO	Executive Order
EU	European Union
MRO	Multistakeholder Regulatory Organization
CAITE	Copyleft AI with Trust Enforcement
RAIL	Responsible AI License
IP	Intellectual Property
LoRA	Low-Rank Adaptation
DEI	Diversity, Equity, and Inclusion

References

1. Roose, K. How ChatGPT Kicked Off an A.I. Arms Race. *The New York Times* **2023**.
2. Ferrag, M.A.; Tihanyi, N.; Debbah, M. From LLM Reasoning to Autonomous AI Agents: A Comprehensive Review, 2025, [arXiv:cs/2504.19678]. https://doi.org/10.48550/arXiv.2504.19678.

 doi:10.20944/preprints202506.0680.v1

3. Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. A Survey on Large Language Model Based Autonomous Agents. *Frontiers of Computer Science* **2024**, *18*, 186345. https://doi.org/10.1007/s11704-024-40231-1.
4. Lee, N.; Cai, Z.; Schwarzschild, A.; Lee, K.; Papailiopoulos, D. Self-Improving Transformers Overcome Easy-to-Hard and Length Generalization Challenges, 2025, [arXiv:cs/2502.01612]. https://doi.org/10.48550/arXiv.2502.01612.
5. Robeyns, M.; Szummer, M.; Aitchison, L. A Self-Improving Coding Agent, 2025, [arXiv:cs/2504.15228]. https://doi.org/10.48550/arXiv.2504.15228.
6. Zhao, A.; Huang, D.; Xu, Q.; Lin, M.; Liu, Y.J.; Huang, G. ExpeL: LLM Agents Are Experiential Learners. *Proceedings of the AAAI Conference on Artificial Intelligence* **2024**, *38*, 19632–19642. https://doi.org/10.1609/aaai.v38i17.29936.
7. Metz, C. A.I. Start-Up Anthropic Challenges OpenAI and Google With New Chatbot. *The New York Times* **2024**.
8. Ostrowski, J. Regulating Machine Learning Open-Source Software: A Primer for Policymakers. Technical report, Abundance Institute, 2024.
9. DeepSeek-AI.; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025, [arXiv:cs/2501.12948]. https://doi.org/10.48550/arXiv.2501.12948.
10. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models, 2023, [arXiv:cs/2302.13971]. https://doi.org/10.48550/arXiv.2302.13971.
11. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B, 2023, [arXiv:cs/2310.06825]. https://doi.org/10.48550/arXiv.2310.06825.
12. Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. Gemma 3 Technical Report, 2025, [arXiv:cs/2503.19786]. https://doi.org/10.48550/arXiv.2503.19786.
13. Abdin, M.; Jacobs, S.A.; Awan, A.A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H.; et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, 2024, [arXiv:cs/2404.14219]. https://doi.org/10.48550/arXiv.2404.14219.
14. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 Technical Report, 2025, [arXiv:cs/2505.09388]. https://doi.org/10.48550/arXiv.2505.09388.
15. Malartic, Q.; Chowdhury, N.R.; Cojocaru, R.; Farooq, M.; Campesan, G.; Djilali, Y.A.D.; Narayan, S.; Singh, A.; Velikanov, M.; Boussaha, B.E.A.; et al. Falcon2-11B Technical Report, 2024, [arXiv:cs/2407.14885]. https://doi.org/10.48550/arXiv.2407.14885.
16. Egashira, K.; Vero, M.; Staab, R.; He, J.; Vechev, M. Exploiting LLM Quantization, 2024, [arXiv:cs/2405.18137]. https://doi.org/10.48550/arXiv.2405.18137.
17. Hooper, C.; Kim, S.; Mohammadzadeh, H.; Mahoney, M.W.; Shao, Y.S.; Keutzer, K.; Gholami, A. KVQuant: Towards 10 Million Context Length LLM Inference with KV Cache Quantization. *Advances in Neural Information Processing Systems* **2024**, *37*, 1270–1303.
18. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models, 2021, [arXiv:cs/2106.09685]. https://doi.org/10.48550/arXiv.2106.09685.
19. Lang, J.; Guo, Z.; Huang, S. A Comprehensive Study on Quantization Techniques for Large Language Models, 2024, [arXiv:cs/2411.02530]. https://doi.org/10.48550/arXiv.2411.02530.
20. Shi, L.; Zhang, H.; Yao, Y.; Li, Z.; Zhao, H. Keep the Cost Down: A Review on Methods to Optimize LLM' s KV-Cache Consumption, 2024, [arXiv:cs/2407.18003]. https://doi.org/10.48550/arXiv.2407.18003.
21. Zhao, Y.; Lin, C.Y.; Zhu, K.; Ye, Z.; Chen, L.; Zheng, S.; Ceze, L.; Krishnamurthy, A.; Chen, T.; Kasikci, B. Atom: Low-Bit Quantization for Efficient and Accurate LLM Serving. *Proceedings of Machine Learning and Systems* **2024**, *6*, 196–209.
22. Dai, D.; Deng, C.; Zhao, C.; Xu, R.X.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; et al. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models, 2024, [arXiv:cs/2401.06066]. https://doi.org/10.48550/arXiv.2401.06066.
23. Fedus, W.; Zoph, B.; Shazeer, N. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, 2022, [arXiv:cs/2101.03961]. https://doi.org/10.48550/arXiv.2101.03961.

 doi:10.20944/preprints202506.0680.v1

24. Jiang, A.Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Hanna, E.B.; Bressand, F.; et al. Mixtral of Experts, 2024, [arXiv:cs/2401.04088]. https://doi.org/10.48550/arXiv.2401.04088.
25. Schroeder, S. Nvidia's Digits Is a Tiny AI Supercomputer for Your Desk. *Mashable* **2025**.
26. Willhoite, P. Why Apple's M4 MacBook Air Is a Milestone for On-Device AI, 2025.
27. Williams, W. Return of the OG? AMD Unveils Radeon AI Pro R9700, Now a Workstation-Class GPU with 32GB GDDR6, 2025.
28. Chiang, W.L.; Zheng, L.; Sheng, Y.; Angelopoulos, A.N.; Li, T.; Li, D.; Zhang, H.; Zhu, B.; Jordan, M.; Gonzalez, J.E.; et al. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference, 2024, [arXiv:cs/2403.04132]. https://doi.org/10.48550/arXiv.2403.04132.
29. Temsah, A.; Alhasan, K.; Altamimi, I.; Jamal, A.; Al-Eyadhy, A.; Malki, K.H.; Temsah, M.H. DeepSeek in Healthcare: Revealing Opportunities and Steering Challenges of a New Open-Source Artificial Intelligence Frontier. *Cureus* **2025**, *17*, e79221. https://doi.org/10.7759/cureus.79221.
30. McIntosh, F.; Murina, S.; Chen, L.; Vargas, H.A.; Becker, A.S. Keeping Private Patient Data off the Cloud: A Comparison of Local LLMs for Anonymizing Radiology Reports. *European Journal of Radiology Artificial Intelligence* **2025**, *2*, 100020. https://doi.org/10.1016/j.ejrai.2025.100020.
31. Montagna, S.; Ferretti, S.; Klopfenstein, L.C.; Ungolo, M.; Pengo, M.F.; Aguzzi, G.; Magnini, M. Privacy-Preserving LLM-based Chatbots for Hypertensive Patient Self-Management. *Smart Health* **2025**, *36*, 100552. https://doi.org/10.1016/j.smhl.2025.100552.
32. Apaydin, K.; Zisgen, Y. Local Large Language Models for Business Process Modeling. In Proceedings of the Process Mining Workshops; Delgado, A.; Slaats, T., Eds., Cham, 2025; pp. 605–609. https://doi.org/10.1007/978-3-031-82225-4_44.
33. Pavsner, M., S. The Attorney's Ethical Obligations When Using AI. *New York Law Journal* **2023**.
34. Tye, J.C. Exploring the Intersections of Privacy and Generative AI: A Dive into Attorney-Client Privilege and ChatGPT. *Jurimetrics* **2024**, *64*, 309.
35. Sakai, K.; Uehara, Y.; Kashihara, S. Implementation and Evaluation of LLM-Based Conversational Systems on a Low-Cost Device. In Proceedings of the 2024 IEEE Global Humanitarian Technology Conference (GHTC), 2024, pp. 392–399. https://doi.org/10.1109/GHTC62424.2024.10771565.
36. Wester, J.; Schrills, T.; Pohl, H.; van Berkel, N. "As an AI Language Model, I Cannot": Investigating LLM Denials of User Requests. In Proceedings of the Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2024; CHI '24, pp. 1–14. https://doi.org/10.1145/3613904.3642135.
37. Sætra, H.S.; Coeckelbergh, M.; Danaher, J. The AI Ethicist's Dilemma: Fighting Big Tech by Supporting Big Tech. *AI and Ethics* **2022**, *2*, 15–27. https://doi.org/10.1007/s43681-021-00123-7.
38. Srivastava, S. Algorithmic Governance and the International Politics of Big Tech. *Perspectives on Politics* **2023**, *21*, 989–1000. https://doi.org/10.1017/S1537592721003145.
39. Verdegem, P. Dismantling AI Capitalism: The Commons as an Alternative to the Power Concentration of Big Tech. *AI & SOCIETY* **2024**, *39*, 727–737. https://doi.org/10.1007/s00146-022-01437-8.
40. Vekaria, Y.; Canino, A.L.; Levitsky, J.; Ciechonski, A.; Callejo, P.; Mandalari, A.M.; Shafiq, Z. Big Help or Big Brother? Auditing Tracking, Profiling, and Personalization in Generative AI Assistants, 2025, [arXiv:cs/2503.16586]. https://doi.org/10.48550/arXiv.2503.16586.
41. Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.M.; Chen, W.; et al. Parameter-Efficient Fine-Tuning of Large-Scale Pre-Trained Language Models. *Nature Machine Intelligence* **2023**, *5*, 220–235. https://doi.org/10.1038/s42256-023-00626-4.
42. Lermen, S.; Rogers-Smith, C.; Ladish, J. LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B, 2024, [arXiv:cs/2310.20624]. https://doi.org/10.48550/arXiv.2310.20624.
43. Candel, A.; McKinney, J.; Singer, P.; Pfeiffer, P.; Jeblick, M.; Lee, C.M.; Conde, M.V. H2O Open Ecosystem for State-of-the-art Large Language Models, 2023, [arXiv:cs/2310.13012]. https://doi.org/10.48550/arXiv.2310.13012.
44. Zhang, D.; Feng, T.; Xue, L.; Wang, Y.; Dong, Y.; Tang, J. Parameter-Efficient Fine-Tuning for Foundation Models, 2025, [arXiv:cs/2501.13787]. https://doi.org/10.48550/arXiv.2501.13787.
45. Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; Ma, Y. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models, 2024, [arXiv:cs/2403.13372]. https://doi.org/10.48550/arXiv.2403.13372.
46. Lyu, K.; Zhao, H.; Gu, X.; Yu, D.; Goyal, A.; Arora, S. Keeping LLMs Aligned After Fine-tuning: The Crucial Role of Prompt Templates, 2025, [arXiv:cs/2402.18540]. https://doi.org/10.48550/arXiv.2402.18540.

 doi:10.20944/preprints202506.0680.v1

47. Nguyen, M.; Baker, A.; Neo, C.; Roush, A.; Kirsch, A.; Shwartz-Ziv, R. Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs, 2025, [arXiv:cs/2407.01082]. https://doi.org/10.48550/arXiv.2407.01082.
48. Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; Jordanous, A. Is Temperature the Creativity Parameter of Large Language Models?, 2024, [arXiv:cs/2405.00492]. https://doi.org/10.48550/arXiv.2405.00492.
49. Goldstein, J.A.; Sastry, G. The Coming Age of AI-Powered Propaganda. *Foreign Affairs* **2023**.
50. Goldstein, J.A.; Chao, J.; Grossman, S.; Stamos, A.; Tomz, M. How Persuasive Is AI-generated Propaganda? *PNAS Nexus* **2024**, *3*, pgae034. https://doi.org/10.1093/pnasnexus/pgae034.
51. Spitale, G.; Biller-Andorno, N.; Germani, F. AI Model GPT-3 (Dis)Informs Us Better than Humans. *Science Advances* **2023**, *9*, eadh1850. https://doi.org/10.1126/sciadv.adh1850.
52. Buchanan, B.; Lohn, A.; Musser, M. Truth, Lies, and Automation. Technical report, Center for Security and Emerging Technology, 2021.
53. Kreps, S.; McCain, R.M.; Brundage, M. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science* **2022**, *9*, 104–117. https://doi.org/10.1017/XPS.2020.37.
54. Wack, M.; Ehrett, C.; Linvill, D.; Warren, P. Generative Propaganda: Evidence of AI's Impact from a State-Backed Disinformation Campaign. *PNAS Nexus* **2025**, *4*, pgaf083. https://doi.org/10.1093/pnasnexus/pgaf083.
55. Thomas, E. "Hey, Fellow Humans!": What Can a ChatGPT Campaign Targeting pro-Ukraine Americans Tell Us about the Future of Generative AI and Disinformation?
56. Barman, D.; Guo, Z.; Conlan, O. The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination. *Machine Learning with Applications* **2024**, *16*, 100545. https://doi.org/10.1016/j.mlwa.2024.100545.
57. Visnjic, D. Generative Models and Deepfake Technology: A Qualitative Research on the Intersection of Social Media and Political Manipulation. In Proceedings of the Artificial Intelligence and Machine Learning; Soliman, K.S., Ed., Cham, 2025; pp. 75–80. https://doi.org/10.1007/978-3-031-79086-7_7.
58. Herbold, S.; Trautsch, A.; Kikteva, Z.; Hautli-Janisz, A. Large Language Models Can Impersonate Politicians and Other Public Figures, 2024, [arXiv:cs/2407.12855]. https://doi.org/10.48550/arXiv.2407.12855.
59. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 Herd of Models, 2024, [arXiv:cs/2407.21783]. https://doi.org/10.48550/arXiv.2407.21783.
60. Williams, A.R.; Burke-Moore, L.; Chan, R.S.Y.; Enock, F.E.; Nanni, F.; Sippy, T.; Chung, Y.L.; Gabasova, E.; Hackenburg, K.; Bright, J. Large Language Models Can Consistently Generate High-Quality Content for Election Disinformation Operations. *PLOS ONE* **2025**, *20*, e0317421. https://doi.org/10.1371/journal.pone.0317421.
61. Haque, M.A. LLMs: A Game-Changer for Software Engineers? *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* **2025**, p. 100204. https://doi.org/10.1016/j.tbench.2025.100204.
62. Idrisov, B.; Schlippe, T. Program Code Generation with Generative AIs. *Algorithms* **2024**, *17*, 62. https://doi.org/10.3390/a17020062.
63. Jiang, J.; Wang, F.; Shen, J.; Kim, S.; Kim, S. A Survey on Large Language Models for Code Generation, 2024, [arXiv:cs/2406.00515]. https://doi.org/10.48550/arXiv.2406.00515.
64. Kirova, V.D.; Ku, C.S.; Laracy, J.R.; Marlowe, T.J. Software Engineering Education Must Adapt and Evolve for an LLM Environment. In Proceedings of the Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1, New York, NY, USA, 2024; SIGCSE 2024, pp. 666–672. https://doi.org/10.1145/3626252.3630927.
65. Coignion, T.; Quinton, C.; Rouvoy, R. A Performance Study of LLM-Generated Code on Leetcode. In Proceedings of the Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering, New York, NY, USA, 2024; EASE '24, pp. 79–89. https://doi.org/10.1145/3661167.3661221.
66. Lebed, S.V.; Namiot, D.E.; Zubareva, E.V.; Khenkin, P.V.; Vorobeva, A.A.; Svichkar, D.A. Large Language Models in Cyberattacks. *Doklady Mathematics* **2024**, *110*, S510–S520. https://doi.org/10.1134/S1064562425700012.
67. Madani, P. Metamorphic Malware Evolution: The Potential and Peril of Large Language Models. In Proceedings of the 2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), 2023, pp. 74–81. https://doi.org/10.1109/TPS-ISA58951.2023.00019.

 doi:10.20944/preprints202506.0680.v1

68. Afane, K.; Wei, W.; Mao, Y.; Farooq, J.; Chen, J. Next-Generation Phishing: How LLM Agents Empower Cyber Attackers. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), 2024, pp. 2558–2567. https://doi.org/10.1109/BigData62323.2024.10825018.
69. Cerullo, M. AI Scams Mimicking Voices Are on the Rise - CBS News. *CBS News* **2023**.
70. Kadali, D.K.; Narayana, K.S.S.; Haritha, P.; Mohan, R.N.V.J.; Kattula, R.; Swamy, K.S.V. Predictive Analysis of Cloned Voice to Commit Cybercrimes Using Generative AI Scammers. In *Algorithms in Advanced Artificial Intelligence*; CRC Press, 2025.
71. Toapanta, F.; Rivadeneira, B.; Tipantuña, C.; Guamán, D. AI-Driven Vishing Attacks: A Practical Approach. *Engineering Proceedings* **2024**, *77*, 15. https://doi.org/10.3390/engproc2024077015.
72. Benegas, G.; Batra, S.S.; Song, Y.S. DNA Language Models Are Powerful Predictors of Genome-Wide Variant Effects. *Proceedings of the National Academy of Sciences* **2023**, *120*, e2311219120. https://doi.org/10.1073/pnas.2311219120.
73. Consens, M.E.; Li, B.; Poetsch, A.R.; Gilbert, S. Genomic Language Models Could Transform Medicine but Not Yet. *npj Digital Medicine* **2025**, *8*, 1–4. https://doi.org/10.1038/s41746-025-01603-4.
74. Ji, Y.; Zhou, Z.; Liu, H.; Davuluri, R.V. DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-language in Genome. *Bioinformatics* **2021**, *37*, 2112–2120. https://doi.org/10.1093/bioinformatics/btab083.
75. Madani, A.; Krause, B.; Greene, E.R.; Subramanian, S.; Mohr, B.P.; Holton, J.M.; Olmos, J.L.; Xiong, C.; Sun, Z.Z.; Socher, R.; et al. Large Language Models Generate Functional Protein Sequences across Diverse Families. *Nature Biotechnology* **2023**, pp. 1–8. https://doi.org/10.1038/s41587-022-01618-2.
76. Nguyen, E.; Poli, M.; Durrant, M.G.; Kang, B.; Katrekar, D.; Li, D.B.; Bartie, L.J.; Thomas, A.W.; King, S.H.; Brixi, G.; et al. Sequence Modeling and Design from Molecular to Genome Scale with Evo. *Science (New York, N.Y.)* **2024**, *386*, eado9336. https://doi.org/10.1126/science.ado9336.
77. James, J.S.; Dai, J.; Chew, W.L.; Cai, Y. The Design and Engineering of Synthetic Genomes. *Nature Reviews Genetics* **2025**, *26*, 298–319. https://doi.org/10.1038/s41576-024-00786-y.
78. Schindler, D.; Dai, J.; Cai, Y. Synthetic Genomics: A New Venture to Dissect Genome Fundamentals and Engineer New Functions. *Current Opinion in Chemical Biology* **2018**, *46*, 56–62. https://doi.org/10.1016/j.cbpa.2018.04.002.
79. Pannu, J.; Bloomfield, D.; MacKnight, R.; Hanke, M.S.; Zhu, A.; Gomes, G.; Cicero, A.; Inglesby, T.V. Dual-Use Capabilities of Concern of Biological AI Models. *PLOS Computational Biology* **2025**, *21*, e1012975. https://doi.org/10.1371/journal.pcbi.1012975.
80. Mackelprang, R.; Adamala, K.P.; Aurand, E.R.; Diggans, J.C.; Ellington, A.D.; Evans, S.W.; Fortman, J.L.C.; Hillson, N.J.; Hinman, A.W.; Isaacs, F.J.; et al. Making Security Viral: Shifting Engineering Biology Culture and Publishing. *ACS Synthetic Biology* **2022**, *11*, 522–527. https://doi.org/10.1021/acssynbio.1c00324.
81. Xie, X.; Lokugamage, K.G.; Zhang, X.; Vu, M.N.; Muruato, A.E.; Menachery, V.D.; Shi, P.Y. Engineering SARS-CoV-2 Using a Reverse Genetic System. *Nature Protocols* **2021**, *16*, 1761–1784. https://doi.org/10.1038/s41596-021-00491-8.
82. Li, J.; Zhao, H.; Zheng, L.; An, W. Advances in Synthetic Biology and Biosafety Governance. *Frontiers in Bioengineering and Biotechnology* **2021**, *9*. https://doi.org/10.3389/fbioe.2021.598087.
83. Zhan, Q.; Fang, R.; Bindu, R.; Gupta, A.; Hashimoto, T.; Kang, D. Removing RLHF Protections in GPT-4 via Fine-Tuning, 2024, [arXiv:cs/2311.05553]. https://doi.org/10.48550/arXiv.2311.05553.
84. Allen, J.G.; Loo, J.; Campoverde, J.L.L. Governing Intelligence: Singapore's Evolving AI Governance Framework. *Cambridge Forum on AI: Law and Governance* **2025**, *1*, e12. https://doi.org/10.1017/cfl.2024.12.
85. NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology (U.S.), Gaithersburg, MD, 2023. https://doi.org/10.6028/NIST.AI.100-1.
86. Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L.A.; Comanescu, R.; Akbulut, C.; Stepleton, T.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; et al. Gaps in the Safety Evaluation of Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* **2024**, *7*, 1200–1217. https://doi.org/10.1609/aies.v7i1.31717.
87. Gault, M. AI Trained on 4Chan Becomes 'Hate Speech Machine'. *Vice* **2022**.
88. Castaño, J.; Martínez-Fernández, S.; Franch, X. Lessons Learned from Mining the Hugging Face Repository. In Proceedings of the Proceedings of the 1st IEEE/ACM International Workshop on Methodological Issues with Empirical Studies in Software Engineering, New York, NY, USA, 2024; WSESE '24, pp. 1–6. https://doi.org/10.1145/3643664.3648204.

 doi:10.20944/preprints202506.0680.v1

89. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing, 2020, [arXiv:cs/1910.03771]. https://doi.org/10.48550/arXiv.1910.03771.
90. Bondarenko, M.; Lushnei, S.; Paniv, Y.; Molchanovsky, O.; Romanyshyn, M.; Filipchuk, Y.; Kiulian, A. Sovereign Large Language Models: Advantages, Strategy and Regulations, 2025, [arXiv:cs/2503.04745]. https://doi.org/10.48550/arXiv.2503.04745.
91. Pomfret, J.; Pang, J.; Pomfret, J.; Pang, J. Exclusive: Chinese Researchers Develop AI Model for Military Use on Back of Meta's Llama. *Reuters* **2024**.
92. Yang, X.; Wang, X.; Zhang, Q.; Petzold, L.; Wang, W.Y.; Zhao, X.; Lin, D. Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models, 2023, [arXiv:cs/2310.02949]. https://doi.org/10.48550/arXiv.2310.02949.
93. Yamin, M.M.; Hashmi, E.; Katt, B. Combining Uncensored and Censored LLMs for Ransomware Generation. In Proceedings of the Web Information Systems Engineering – WISE 2024; Barhamgi, M.; Wang, H.; Wang, X., Eds., Singapore, 2025; pp. 189–202. https://doi.org/10.1007/978-981-96-0573-6_14.
94. Wan, A.; Wallace, E.; Shen, S.; Klein, D. Poisoning Language Models During Instruction Tuning, 2023, [arXiv:cs/2305.00944]. https://doi.org/10.48550/arXiv.2305.00944.
95. Barclay, I.; Preece, A.; Taylor, I. Defining the Collective Intelligence Supply Chain, 2018, [arXiv:cs/1809.09444]. https://doi.org/10.48550/arXiv.1809.09444.
96. Hopkins, A.; Cen, S.H.; Ilyas, A.; Struckman, I.; Videgaray, L.; Mądry, A. AI Supply Chains: An Emerging Ecosystem of AI Actors, Products, and Services, 2025, [arXiv:cs/2504.20185]. https://doi.org/10.48550/arXiv.2504.20185.
97. Gstrein, O.J.; Haleem, N.; Zwitter, A. General-Purpose AI Regulation and the European Union AI Act. *Internet Policy Review* **2024**, *13*, 1–26. https://doi.org/10.14763/2024.3.1790.
98. Evas, T. The EU Artificial Intelligence Act:. *Journal of AI Law and Regulation* **2024**, *1*, 98–101. https://doi.org/10.21552/aire/2024/1/11.
99. El Ali, A.; Venkatraj, K.P.; Morosoli, S.; Naudts, L.; Helberger, N.; Cesar, P. Transparent AI Disclosure Obligations: Who, What, When, Where, Why, How. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2024; CHI EA '24, pp. 1–11. https://doi.org/10.1145/3613905.3650750.
100. President, T. Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 2023.
101. Lubello, V. From Biden to Trump: Divergent and Convergent Policies in The Artificial Intelligence (AI) Summer. *DPCE Online* **2025**, *69*. https://doi.org/10.57660/dpceonline.2025.2463.
102. House, T.W. Removing Barriers to American Leadership in Artificial Intelligence. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/, 2025.
103. Franks, E.; Lee, B.; Xu, H. Report: China's New AI Regulations. *Global Privacy Law Review* **2024**, *5*.
104. Diallo, K.; Smith, J.; Okolo, C.T.; Nyamwaya, D.; Kgomo, J.; Ngamita, R. Case Studies of AI Policy Development in Africa. *Data & Policy* **2025**, *7*, e15. https://doi.org/10.1017/dap.2024.71.
105. Dua, M.; Singh, J.P.; Shehu, A. The Ethics of National Artificial Intelligence Plans: An Empirical Lens. *AI and Ethics* **2025**. https://doi.org/10.1007/s43681-025-00663-2.
106. Kulothungan, V.; Gupta, D. Towards Adaptive AI Governance: Comparative Insights from the U.S., EU, and Asia, 2025, [arXiv:cs/2504.00652]. https://doi.org/10.48550/arXiv.2504.00652.
107. Munn, L. The Uselessness of AI Ethics. *AI and Ethics* **2023**, *3*, 869–877. https://doi.org/10.1007/s43681-022-00209-w.
108. O'Brien, M. Tech Industry Tried Reducing AI's Pervasive Bias. Now Trump Wants to End Its 'woke AI' Efforts. *AP News* **2025**.
109. O'Brien, M.; Parvini, S. Trump Signs Executive Order on Developing Artificial Intelligence 'Free from Ideological Bias'. *AP News* **2025**.
110. Mittelstadt, B. Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence* **2019**, *1*, 501–507. https://doi.org/10.1038/s42256-019-0114-4.
111. Quan, E. Censorship Sensing: The Capabilities and Implications of China's Great Firewall Under Xi Jinping. *Sigma: Journal of Political and International Studies* **2022**, *39*, 19–31.
112. Wong, H. Mapping the Open-Source AI Debate: Cybersecurity Implications and Policy Priorities, 2025.
113. Abdelnabi, S.; Fritz, M. Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding. In Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP), 2021, pp. 121–140. https://doi.org/10.1109/SP40001.2021.00083.

 doi:10.20944/preprints202506.0680.v1

114. Uddin, M.S.; Ohidujjaman.; Hasan, M.; Shimamura, T. Audio Watermarking: A Comprehensive Review. *International Journal of Advanced Computer Science and Applications (ijacsa)* **2024**, *15*. https://doi.org/10.14569/IJACSA.2024.01505141.
115. Zhao, X.; Zhang, K.; Su, Z.; Vasan, S.; Grishchenko, I.; Kruegel, C.; Vigna, G.; Wang, Y.X.; Li, L. Invisible Image Watermarks Are Provably Removable Using Generative AI, 2024, [arXiv:cs/2306.01953]. https://doi.org/10.48550/arXiv.2306.01953.
116. Han, T.A.; Lenaerts, T.; Santos, F.C.; Pereira, L.M. Voluntary Safety Commitments Provide an Escape from Over-Regulation in AI Development. *Technology in Society* **2022**, *68*, 101843. https://doi.org/10.1016/j.techsoc.2021.101843.
117. Ali, S.J.; Christin, A.; Smart, A.; Katila, R. Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk among Ethics Entrepreneurs. In Proceedings of the 2023 ACM Conference on Fairness Accountability and Transparency, 2023, pp. 217–226, [arXiv:cs/2305.09573]. https://doi.org/10.1145/3593013.3593990.
118. Varanasi, R.A.; Goyal, N. "It Is Currently Hodgepodge": Examining AI/ML Practitioners' Challenges during Co-production of Responsible AI Values. In Proceedings of the Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2023; CHI '23, pp. 1–17. https://doi.org/10.1145/3544548.3580903.
119. Widder, D.G.; Zhen, D.; Dabbish, L.; Herbsleb, J. It's about Power: What Ethical Concerns Do Software Engineers Have, and What Do They (Feel They Can) Do about Them? In Proceedings of the Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 2023; FAccT '23, pp. 467–479. https://doi.org/10.1145/3593013.3594012.
120. van Maanen, G. AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics. *Digital Society* **2022**, *1*, 9. https://doi.org/10.1007/s44206-022-00013-3.
121. Ferrandis, C.M.; Lizarralde, M.D. Open Sourcing AI: Intellectual Property at the Service of Platform Leadership. *JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law* **2022**, *13*, 224–246. https://doi.org/https://nbn-resolving.de/urn:nbn:de:0009-29-55579.
122. Contractor, D.; McDuff, D.; Haines, J.K.; Lee, J.; Hines, C.; Hecht, B.; Vincent, N.; Li, H. Behavioral Use Licensing for Responsible AI. In Proceedings of the Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 2022; FAccT '22, pp. 778–788. https://doi.org/10.1145/3531146.3533143.
123. Klyman, K. Acceptable Use Policies for Foundation Models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* **2024**, *7*, 752–767. https://doi.org/10.1609/aies.v7i1.31677.
124. McDuff, D.; Korjakow, T.; Cambo, S.; Benjamin, J.J.; Lee, J.; Jernite, Y.; Ferrandis, C.M.; Gokaslan, A.; Tarkowski, A.; Lindley, J.; et al. On the Standardization of Behavioral Use Clauses and Their Adoption for Responsible Licensing of AI, 2024, [arXiv:cs/2402.05979]. https://doi.org/10.48550/arXiv.2402.05979.
125. Schmit, C.D.; Doerr, M.J.; Wagner, J.K. Leveraging IP for AI Governance. *Science* **2023**, *379*, 646–648. https://doi.org/10.1126/science.add2202.
126. Henderson, P.; Lemley, M.A. The Mirage of Artificial Intelligence Terms of Use Restrictions, 2024, [arXiv:cs/2412.07066]. https://doi.org/10.48550/arXiv.2412.07066.
127. Cui, J.; Araujo, D.A. Rethinking Use-Restricted Open-Source Licenses for Regulating Abuse of Generative Models. *Big Data & Society* **2024**, *11*, 20539517241229699. https://doi.org/10.1177/20539517241229699.
128. Crouch, D. Using Intellectual Property to Regulate Artificial Intelligence. *Missouri Law Review* **2024**, *89*, 781.
129. Widder, D.G.; Nafus, D.; Dabbish, L.; Herbsleb, J. Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. In Proceedings of the Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 2022; FAccT '22, pp. 2035–2046. https://doi.org/10.1145/3531146.3533779.
130. Pawelec, M. Decent Deepfakes? Professional Deepfake Developers' Ethical Considerations and Their Governance Potential. *AI and Ethics* **2024**. https://doi.org/10.1007/s43681-024-00542-2.
131. Maktabdar Oghaz, M.; Babu Saheer, L.; Dhame, K.; Singaram, G. Detection and Classification of ChatGPT-generated Content Using Deep Transformer Models. *Frontiers in Artificial Intelligence* **2025**, *8*, 1458707. https://doi.org/10.3389/frai.2025.1458707.
132. Rashidi, H.H.; Fennell, B.D.; Albahra, S.; Hu, B.; Gorbett, T. The ChatGPT Conundrum: Human-generated Scientific Manuscripts Misidentified as AI Creations by AI Text Detection Tool. *Journal of Pathology Informatics* **2023**, *14*, 100342. https://doi.org/10.1016/j.jpi.2023.100342.

 doi:10.20944/preprints202506.0680.v1

133. Weber-Wulff, D.; Anohina-Naumeca, A.; Bjelobaba, S.; Foltýnek, T.; Guerrero-Dib, J.; Popoola, O.; Šigut, P.; Waddington, L. Testing of Detection Tools for AI-generated Text. *International Journal for Educational Integrity* **2023**, *19*, 26. https://doi.org/10.1007/s40979-023-00146-z.
134. Poireault, K. Malicious AI Models on Hugging Face Exploit Novel Attack Technique. https://www.infosecurity-magazine.com/news/malicious-ai-models-hugging-face/, 2025.
135. Sabt, M.; Achemlal, M.; Bouabdallah, A. Trusted Execution Environment: What It Is, and What It Is Not. In Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA, 2015, Vol. 1, pp. 57–64. https://doi.org/10.1109/Trustcom.2015.357.
136. AĞCA, M.A.; Faye, S.; Khadraoui, D. A Survey on Trusted Distributed Artificial Intelligence. *IEEE Access* **2022**, *10*, 55308–55337. https://doi.org/10.1109/ACCESS.2022.3176385.
137. Geppert, T.; Deml, S.; Sturzenegger, D.; Ebert, N. Trusted Execution Environments: Applications and Organizational Challenges. *Frontiers in Computer Science* **2022**, *4*. https://doi.org/10.3389/fcomp.2022.930741.
138. Jauernig, P.; Sadeghi, A.R.; Stapf, E. Trusted Execution Environments: Properties, Applications, and Challenges. *IEEE Security & Privacy* **2020**, *18*, 56–60. https://doi.org/10.1109/MSEC.2019.2947124.
139. Babar, M.F.; Hasan, M. Trusted Deep Neural Execution—A Survey. *IEEE Access* **2023**, *11*, 45736–45748. https://doi.org/10.1109/ACCESS.2023.3274190.
140. Cai, Z.; Ma, R.; Fu, Y.; Zhang, W.; Ma, R.; Guan, H. LLMaaS: Serving Large-Language Models on Trusted Serverless Computing Platforms. *IEEE Transactions on Artificial Intelligence* **2025**, *6*, 405–415. https://doi.org/10.1109/TAI.2024.3429480.
141. Dong, B.; Wang, Q. Evaluating the Performance of the DeepSeek Model in Confidential Computing Environment, 2025, [arXiv:cs/2502.11347]. https://doi.org/10.48550/arXiv.2502.11347.
142. Greamo, C.; Ghosh, A. Sandboxing and Virtualization: Modern Tools for Combating Malware. *IEEE Security & Privacy* **2011**, *9*, 79–82. https://doi.org/10.1109/MSP.2011.36.
143. Prevelakis, V.; Spinellis, D. Sandboxing Applications. In Proceedings of the USENIX Annual Technical Conference, FREENIX Track, 2001, pp. 119–126.
144. Johnson, J. The AI Commander Problem: Ethical, Political, and Psychological Dilemmas of Human-Machine Interactions in AI-enabled Warfare. *Journal of Military Ethics* **2022**, *21*, 246–271. https://doi.org/10.1080/15027570.2023.2175887.
145. Salo-Pöntinen, H. AI Ethics - Critical Reflections on Embedding Ethical Frameworks in AI Technology. In Proceedings of the Culture and Computing. Design Thinking and Cultural Computing; Rauterberg, M., Ed., Cham, 2021; pp. 311–329. https://doi.org/10.1007/978-3-030-77431-8_20.
146. Cai, Y.; Liang, P.; Wang, Y.; Li, Z.; Shahin, M. Demystifying Issues, Causes and Solutions in LLM Open-Source Projects. *Journal of Systems and Software* **2025**, *227*, 112452. https://doi.org/10.1016/j.jss.2025.112452.
147. Win, H.M.; Wang, H.; Tan, S.H. Towards Automated Detection of Unethical Behavior in Open-Source Software Projects. In Proceedings of the Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, New York, NY, USA, 2023; ESEC/FSE 2023, pp. 644–656. https://doi.org/10.1145/3611643.3616314.
148. Wang, W. Rethinking AI Safety Approach in the Era of Open-Source AI, 2025.
149. Carlisle, K.; Gruby, R.L. Polycentric Systems of Governance: A Theoretical Model for the Commons. *Policy Studies Journal* **2019**, *47*, 927–952. https://doi.org/10.1111/psj.12212.
150. Ostrom, E. Polycentric Systems for Coping with Collective Action and Global Environmental Change. *Global Environmental Change* **2010**, *20*, 550–557. https://doi.org/10.1016/j.gloenvcha.2010.07.004.
151. Huang, L.T.L.; Papyshev, G.; Wong, J.K. Democratizing Value Alignment: From Authoritarian to Democratic AI Ethics. *AI and Ethics* **2025**, *5*, 11–18. https://doi.org/10.1007/s43681-024-00624-1.
152. Cihon, P.; Maas, M.M.; Kemp, L. Should Artificial Intelligence Governance Be Centralised? Design Lessons from History. In Proceedings of the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 2020; AIES '20, pp. 228–234. https://doi.org/10.1145/3375627.3375857.
153. Attard-Frost, B.; Widder, D.G. The Ethics of AI Value Chains. *Big Data & Society* **2025**, *12*, 20539517251340603. https://doi.org/10.1177/20539517251340603.
154. Muldoon, J.; Cant, C.; Graham, M.; Ustek Spilda, F. The Poverty of Ethical AI: Impact Sourcing and AI Supply Chains. *AI & SOCIETY* **2025**, *40*, 529–543. https://doi.org/10.1007/s00146-023-01824-9.
155. Widder, D.G.; Nafus, D. Dislocated Accountabilities in the "AI Supply Chain": Modularity and Developers' Notions of Responsibility. *Big Data & Society* **2023**, *10*, 20539517231177620. https://doi.org/10.1177/20539517231177620.

 doi:10.20944/preprints202506.0680.v1

156. McKelvey, F.; MacDonald, M. Artificial Intelligence Policy Innovations at the Canadian Federal Government. *Canadian Journal of Communication* **2019**, *44*, PP–43. https://doi.org/10.22230/cjc.2019v44n2a3509.
157. Stahl, B.C.; Antoniou, J.; Bhalla, N.; Brooks, L.; Jansen, P.; Lindqvist, B.; Kirichenko, A.; Marchal, S.; Rodrigues, R.; Santiago, N.; et al. A Systematic Review of Artificial Intelligence Impact Assessments. *Artificial Intelligence Review* **2023**, *56*, 12799–12831. https://doi.org/10.1007/s10462-023-10420-8.
158. Hsu, Y.C.; Huang, T.H.K.; Verma, H.; Mauri, A.; Nourbakhsh, I.; Bozzon, A. Empowering Local Communities Using Artificial Intelligence. *Patterns* **2022**, *3*. https://doi.org/10.1016/j.patter.2022.100449.
159. Esteves, A.M.; , Daniel, F.; and Vanclay, F. Social Impact Assessment: The State of the Art. *Impact Assessment and Project Appraisal* **2012**, *30*, 34–42. https://doi.org/10.1080/14615517.2012.660356.
160. Welsh, C.; Román García, S.; Barnett, G.C.; Jena, R. Democratising Artificial Intelligence in Healthcare: Community-Driven Approaches for Ethical Solutions. *Future Healthcare Journal* **2024**, *11*, 100165. https://doi.org/10.1016/j.fhj.2024.100165.
161. Agnese, P.; Arduino, F.R.; Prisco, D.D. The Era of Artificial Intelligence: What Implications for the Board of Directors? *Corporate Governance: The International Journal of Business in Society* **2024**, *25*, 272–287. https://doi.org/10.1108/CG-06-2023-0259.
162. Collina, L.; Sayyadi, M.; Provitera, M. Critical Issues About A.I. Accountability Answered. *California Management Review Insights* **2023**.
163. da Fonseca, A.T.; Vaz de Sequeira, E.; Barreto Xavier, L. Liability for AI Driven Systems. In *Multidisciplinary Perspectives on Artificial Intelligence and the Law*; Sousa Antunes, H.; Freitas, P.M.; Oliveira, A.L.; Martins Pereira, C.; Vaz de Sequeira, E.; Barreto Xavier, L., Eds.; Springer International Publishing: Cham, 2024; pp. 299–317. https://doi.org/10.1007/978-3-031-41264-6_16.
164. Buiten, M.; de Streel, A.; Peitz, M. The Law and Economics of AI Liability. *Computer Law & Security Review* **2023**, *48*, 105794. https://doi.org/10.1016/j.clsr.2023.105794.
165. Ramakrishnan, K.; Smith, G.; Downey, C. U.S. Tort Liability for Large-Scale Artificial Intelligence Damages: A Primer for Developers and Policymakers. Technical report, Rand Corporation, 2024.
166. Andrews, C. European Commission Withdraws AI Liability Directive from Consideration, 2025.
167. Tschider, C. Will a Cybersecurity Safe Harbor Raise All Boats? *Lawfare* **3/20/2024 1:42:01 PM**.
168. Shinkle, D. The Ohio Data Protection Act: An Analysis of the Ohio Cybersecurity Safe Harbor. *University of Cincinnati Law Review* **2019**, *87*, 1213–1235.
169. Oberly, D.J. A Potential Trend in the Making? Utah Becomes the Second State to Enact Data Breach Safe Harbor Law Incentivizing Companies to Maintain Robust Data Protection Programs. *ABA TIPS Cybersecurity & Data Privacy Committee Newsletter* **2021**.
170. Blumstein, J.F.; McMichael, B.J.; Storrow, A.B. Developing Safe Harbors to Address Malpractice Liability and Wasteful Health Care Spending. *JAMA Health Forum* **2023**, *4*, e233899. https://doi.org/10.1001/jamahealthforum.2023.3899.
171. Dove, E.S.; Knoppers, B.M.; Zawati, M.H. Towards an Ethics Safe Harbor for Global Biomedical Research. *Journal of Law and the Biosciences* **2014**, *1*, 3–51. https://doi.org/10.1093/jlb/lst002.
172. McNerney, J. McNerney Introduces Bill to Establish Safety Standards for Artificial Intelligence While Fostering Innovation, 2025.
173. Just, N.; Latzer, M. Governance by Algorithms: Reality Construction by Algorithmic Selection on the Internet. *Media, Culture & Society* **2017**, *39*, 238–258. https://doi.org/10.1177/0163443716643157.
174. Khanal, S.; Zhang, H.; Taeihagh, A. Why and How Is the Power of Big Tech Increasing in the Policy Process? The Case of Generative AI. *Policy and Society* **2025**, *44*, 52–69. https://doi.org/10.1093/polsoc/puae012.