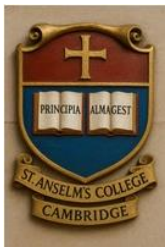

Journal of Classical Pure Mathematics – St. Anselm's College, Cambridge, UK

Aarav Kulshrestha*
Primary Contact:
adiworks129@gmail.com

Towards Ethical AI: Policy Proposals for Addressing Bias, Security, Copyright, and Liability Concerns in Artificial Intelligence

Prepared by:
Aarav Kulshrestha

Under the mentorship of:
Dr Jonathan Kenigson

For:
AI Journal, London

Abstract.

Artificial intelligence (AI) is transforming industries, governance, and daily life, but its rapid adoption raises significant ethical and regulatory challenges. This paper explores the intersections of bias, copyright, liability, and security in AI systems, analyzing real-world cases such as generative art copyright disputes, algorithmic bias in chatbots, and AI-assisted military technologies. Using qualitative policy analysis and case study methodology, the research identifies structural gaps in existing legal and regulatory frameworks that leave creators, users, and society vulnerable. Findings highlight the need for standardized yet flexible approaches, including compulsory licensing for copyrighted data, algorithmic impact assessments, tiered liability frameworks, and dual-use security regulations. By proposing integrated and globally coordinated policies, the thesis demonstrates how responsible AI development can balance innovation with accountability, equity, and security, providing a blueprint for ethical AI governance.

“A tank that drives itself, a drone that picks its own targets; and a machine gun with facial recognition software. Weapons powered by AI are already here” (“A.I. Kill”). These technologies demonstrate the alacritous integration of artificial intelligence into military and security systems, raising concerns about control, accountability, and ethical implications. Beyond the battlefield, similar advancements are transforming civilian industries, introducing both new opportunities and regulatory challenges (Schönfuß). Policymakers are working to develop a comprehensive understanding of AI across its various applications—including engineering, healthcare, and military use—to effectively address its challenges while still supporting its growth. As AI technology advances, policies and discussions are aimed at addressing copyright, bias, and liability issues while promoting innovation. However, inconsistencies in policies across jurisdictions highlight the need for a standardized yet adaptable framework to ensure AI's responsible development. This paper proposes a unified, globally applicable framework for AI policy adoption.

Artificial Intelligence (A.I.) is notorious for violating copyright claims through the absorption of data. AI tools process vast amounts of data, mimicking human learning at an accelerated pace. Issues arise when the A.I. produces output strikingly like the input data. This perceived infringement has sparked concerns among artists regarding copyright violations. Consider Kelly McKernan, a visual artist shaped by these experiences, who in adolescence shared her paintings in an art gallery called DeviantArt, purely motivated by devotion to her craft.

DeviantArt introduced a subscription-based AI art generator. And this AI art generator had been trained on countless images from artists like Kelly, but the DeviantArt artists

would not get a cent. After hearing about this, Kelly and a group of other DeviantArt artists got together and created a class-action lawsuit (“Artists”). This lawsuit is a real-world example of A.I infringing on copyright without proper permission. The absence of clear policies puts both companies and artists at risk. Andres Guadamuz, a Senior Lecturer in intellectual property law at the University of Sussex, characterizes A.I. models as entities that learn patterns from original images, brushstrokes, and artistic styles. He points out that these elements fall outside the scope of copyright law. However, Andres still believes that copyright could still pose an issue (“Artists”). As AI evolves and stores more information, the problem worsens daily. This lack of A.I. related copyright policy underscores the urgent need for regulations to prevent AI from creating irreparable harm. Deviantart is not the only company under attack; Getty Images is inculpating the creators of Stability AI for stealing its content. The A.I. company is being accused of the “brazen infringement of Getty Images intellectual property on a staggering scale.” Getty claims that Stability AI copied more than 12 million images from its database “without permission ... or compensation ... as part of its efforts to build a competing business,” and that the startup has infringed on both the company’s copyright and trademark protections (Vincent). One suggested solution for the problem is that copyright owners may be able to show that such outputs generated by AI infringe their copyrights if the AI program both had access to their copyrighted work(s) and generated substantially similar outputs (Zirpoli). This, however, is just one solution to a complex issue. Luckily, companies like OpenAI, Google, and others are now watermarking AI content to help fight this issue of copyright infringement (Bartz). Additionally, Spain approved a law imposing fines of up to €35 million or 7% of annual

global turnover on companies that fail to label AI-generated content, aiming to combat the spread of deepfakes and protect consumers from AI-generated disinformation (Desmarais). We can also look to the European Union (EU) A.I. Act, which requires A.I. systems to disclose copyrighted material during training and mandates labeling of A.I. content (Guadamuz). These, however, are not widely implemented solutions, and hence they fail to provide solvency for any problems. Rebecca Kelly Slaughter, the commissioner of the United States Federal Trade Commission, builds from this point; she explains how the lack of protection when it comes to AI copyright infringement has impacted some of the most important areas of the American economy (Slaughter). The lack of proper and organized policy is being directly linked with economic drawbacks, as explained by the commissioner.

Now to the proposed policies: The cases of Kelly McKernan and Getty Images illustrate a clear clash between owners of copyrighted works and the AI companies using those works. The problem is strikingly apparent when creators have to rely on costly litigation for a problem that should be safeguarded *ex ante* through *statutory* design. A more effective policy should mandate the creation of compulsory data licensing frameworks, akin to those in the music and broadcasting industries. Systems should require dataset holders to pay for copyrighted material, using a centralized platform to track and distribute payments in accordance with usage instead of forcing individual artists to sue. In this apparatus, companies possessing copyrighted works in training must contribute to a centralized rights management pool overseen by an independent authority. Current dataset holders extract data without paying creators, relying on claims of ‘fair use’ or claiming exemption under data training purposes.

Centralized licensing solves this problem as it eliminates the need for individual negotiations, reduces judicial congestion and costly litigation, ensures fair compensation for artists, and removes the need to prove 'substantial similarity' in each case – a point often fiercely debated in AI copyright. Governments should require dataset provenance tools – such as blockchain, cryptographic hashing, or machine-readable watermarking – to document and verify data sourcing. Additionally, by holding dataset holders accountable rather than every AI developer, small developers using preexisting datasets remain unaffected, while large companies with proprietary datasets must comply with licensing and provenance rules. These records would not only facilitate compensation but also provide courts with evidentiary clarity in infringement disputes. Were the disclosure requirements outlined in the European Union's AI Act – e.g. compulsory documentation of datasets and the labeling of AI-generated outputs – to be implemented, they would create consistency across jurisdictions and address the interoperability issues that currently hinder the AI landscape. Only through such mechanisms can the pernicious dynamic identified in your analysis – the uncompensated appropriation of artistic labor – be structurally curtailed rather than episodically litigated.

The proposed centralized licensing and provenance systems address the core issues shown by the cases of Kelly McKernan and Getty Images, but several practical challenges still persist. For one, without verifiable records of dataset origins, enforcing licensing obligations is technically difficult, allowing copyrighted works to be used without proper compensation. To overcome these issues, AI developers should adopt cryptographically verifiable provenance tools that provide clear, tamper-proof

documentation of all training data, directly supporting the rights management framework described above. Additionally, tiered licensing fees and sandbox exemptions (where costs scale with the size or resources of the use) would allow smaller actors to participate without undue burden. This can be paired with automated monitoring systems that automatically check datasets for proper licensing before use, reducing copyright violations and minimizing the need for costly litigation by detecting unlicensed content before it is deployed, reducing the need for costly litigation and ensuring that innovation and creator rights coexist sustainably. Sandboxes could be like the United Kingdom's regulatory sandbox, which serves as a regulatory ideal for helping the fintech and crypto industries already. Sandboxes create experimental environments that let small developers, startups, or researchers test AI systems with reduced regulatory burden, encouraging innovation without excessive costs. Combining provenance tools, tiered licensing, sandbox exemptions, and automated monitoring can protect creators' rights while allowing innovation to continue.

However, even with centralized licensing, the system relies on complete and accurate reporting of all copyrighted works. If dataset holders fail to report certain content, or if transformed or scraped works evade detection, creators may not receive proper compensation. Some critics also argue that centralized licensing and provenance requirements could stifle innovation or create excessive bureaucratic overhead, particularly for smaller developers. To mitigate these, we can retrace to previous ideas: automated monitoring tools, tiered enforcement with heavier obligations for large dataset holders, sandbox exemptions and international coordination of standards can help close potential loopholes and ensure the framework remains

effective and fair. Automated monitoring tools can detect unreported or improperly transformed content in datasets, helping ensure creators receive proper compensation even when some works might otherwise evade detection. Tiered enforcement with heavier obligations for large dataset holders ensures that major actors comply with licensing requirements, reducing loopholes that could undermine the framework's fairness. Sandbox exemptions allow small developers to innovate without undue burden, while international coordination of standards ensures consistent enforcement across jurisdictions, preventing gaps that could let unlicensed content go unaddressed. This design demonstrates that legal safeguards and technical feasibility can coexist, addressing both creator rights and the broader interests of the AI development community.

While blockchain registries, cryptographic hashing, and digital watermarking offer promising methods to track and verify the provenance of AI training data, each faces practical limitations that must be addressed to be effective. Blockchain provides a permanent, tamper-proof record of dataset inputs, but maintaining such ledgers for millions or billions of files can be computationally expensive, slow, and difficult to scale. This could be mitigated by using layered or sharded blockchain architectures that distribute the computational load, combined with selective recording of only key fingerprints or high value works rather than the entire dataset. Cryptographic hashes uniquely identify files, yet even minor alterations can produce entirely new hashes, allowing slightly modified works to bypass detection. One potential solution is fuzzy or perceptual hashing, which generates fingerprints that remain consistent across minor modifications, making it easier to detect derivative or transformed works. Watermarks

embed ownership information directly into files, but they may be removed, corrupted, or degraded during preprocessing or AI training transformations. To address this, robust, multi-layer watermarking techniques could be used, embedding redundant and imperceptible marks that survive common transformations, alongside machine-learning-based detection algorithms that can identify degraded or partially altered watermarks. By combining these approaches, it is possible to create a resilient, verifiable provenance system that balances technical feasibility with legal enforceability, ensuring that creators' rights are protected while minimizing burden on dataset holders and AI developers.

In response to the challenges faced by creators like Kelly McKernan and Getty Images, governments can implement clear policies to ensure AI respects copyright while supporting innovation. Centralized licensing frameworks should require dataset holders to contribute to a rights management pool, with payments automatically distributed to creators. Verifiable provenance tools – such as blockchain registries, cryptographic hashing, and robust digital watermarking – should be mandated, with flexibility for innovations like fuzzy hashing or layered watermarks. Regulators could audit datasets and treat missing or incomplete provenance as evidence of infringement. Small developers and academic projects should be exempt, focusing obligations on large proprietary dataset holders. Finally, aligning domestic rules with international standards, like the EU AI Act, can prevent circumvention across borders. Together, these measures create a legally enforceable, technically feasible system that protects creators, ensures fair compensation, and addresses the economic and creative risks highlighted by these landmark cases.

Artificial Intelligence has the potential to be biased in its actions. The problem occurs when an A.I. system learns from biased sources. Shown in the article, “Managing Bias in A.I.”, prejudiced A.I. is an issue because people use A.I. for its unbiased approach to find solutions to their problems. The article goes on to explain how the rising awareness of bias is putting A.I. companies at risk of losing the trust of their clients. The companies face an even harder time trying to weed out discriminatory features, due to the fact that A.I. algorithms are mostly inexplicable compared to other code (Roselli, Drew et al.). This is because “[e]ven with careful review of the algorithms and data sets, it may not be possible to delete all unwanted biases, particularly because A.I. systems learn from historical data, which encodes historical biases” (Roselli, Drew et al.). Take for example, Perspective API, initially released by Alphabet’s Jigsaw in 2017, is a tool for measuring toxicity in natural language. The possibility of bias created from this program is almost endless, as it is highly opinionated on what is toxic or not. Just in 2022 alone Perspective A.I. was found to be used in 37 research papers (“Index Report”). The Perspective API, created to assess toxicity in natural language, demonstrates the potential for bias as it relies on an inherently subjective framework to define what constitutes “toxic” behavior, leading to inconsistencies in its application and potentially reinforcing harmful stereotypes or misunderstandings, which has been widely used in academic research to study its limitations and effectiveness in 37 different studies in 2022 alone. There is even gender representation bias in chatbots; A study found that 37% of bots are female, 20% are male, 3% are both, and 40% are genderless (“Gender Representation”). The finding that 37% of chatbots are designed with female personas, 20% with male personas, and 40% being labeled as genderless,

reveals a critical bias in the way AI developers are shaping gender representation, which can perpetuate societal stereotypes, limit inclusivity, and influence the way users interact with these systems, particularly when a majority of chatbots are non-gendered or reflect gender biases.

The second axis of reform concerns algorithmic bias, a phenomenon that, as the Perspective API case study demonstrates, corrodes public trust precisely because it undermines the expectation of neutrality. Current voluntary guidelines issued by corporations are inadequate, as they lack both enforceability and comparability. It is critical to note that the objective of these proposals is not to eradicate all AI bias, as certain biases merely reflect historical or societal patterns and are not per se unlawful. Rather, the policy's focus is on mitigating biases that produce legally actionable harms, particularly those that result in disparate treatment or discrimination in legally protected domains. From a U.S. perspective, the Equal Credit Opportunity Act (ECOA), Title VII of the Civil Rights Act, and the Americans with Disabilities Act (ADA) establish obligations against discrimination in credit, employment, and access to services; algorithmic systems that produce outputs violating these statutes would create clear legal exposure for developers and deploying entities. Similarly, in the European Union, the General Data Protection Regulation (GDPR) imposes obligations regarding automated decision-making and profiling, requiring transparency, meaningful human oversight, and safeguards against discriminatory processing (Articles 22 and 5(1)(a)). Moreover, the proposed EU AI Act explicitly classifies high-risk AI systems—those affecting employment, education, law enforcement, and social services—as subject to mandatory conformity assessments, risk management procedures, and bias mitigation measures,

establishing enforceable standards for equitable AI deployment. Against this legal backdrop, focusing policy interventions on harmful or discriminatory biases ensures alignment with both civil rights law and data protection frameworks, providing a defensible statutory basis for intervention without overreaching into neutral or contextually permissible biases.

Now to address algorithmic bias, a phenomenon that, as the Perspective API case demonstrates, is at the core of public trust by violating the expectation of neutrality in automated decision-making. Existing voluntary corporate guidelines are insufficient because they lack enforceability, consistency, and comparability across platforms, leaving high-risk AI systems susceptible to discriminatory outputs. While bias in AI often reflects historical social patterns and is not inherently unlawful, it becomes legally actionable when it produces outcomes that contravene anti-discrimination statutes, including Title VII of the U.S. Civil Rights Act, the Equality Act 2010 (UK), and Articles 21 and 23 of the EU Charter of Fundamental Rights, as well as obligations under the EU AI Act for high-risk systems. To mitigate these harmful effects, AI deployed in sensitive domains – such as employment, credit scoring, criminal justice risk assessment, education, and healthcare – should undergo statutory Algorithmic Impact Assessments (AIAs), requiring disclosure of training data demographics, documentation of bias mitigation strategies, application of fairness metrics (e.g., demographic parity, equalized odds, disparate impact ratios), independent third-party audits, and explainability reports for outputs with material social, economic, or legal consequences, in alignment with GDPR Article 22 and EU AI Act transparency standards. To ensure technical rigor, governments should establish bias benchmarking consortia, analogous

to the NIST in cybersecurity, producing standardized evaluation suites, defining acceptable bias thresholds, and publishing public benchmarks to guide compliance and remediation, complemented by continuous monitoring pipelines, adversarial testing, and automated detection of emergent biases. Regulatory obligations should be tiered, applying stringent requirements to large-scale proprietary systems while reducing burdens for small developers, academic researchers, or experimental deployments using preexisting licensed datasets, with periodic audits and corrective mechanisms mandated to address detected disparities. By integrating AIAs, standardized fairness metrics, explainability reporting, auditing, monitoring, benchmarking consortia, and tiered obligations, this framework provides a legally enforceable and technically robust strategy to mitigate discriminatory AI outcomes, uphold anti-discrimination obligations, and preserve opportunities for responsible AI innovation across jurisdictions.

Algorithmic Impact Assessments (AIAs), disclosure of training data demographics, and fairness-metric mandates are conceptually powerful. However, they often provoke three linked critiques, which seasoned practitioners have observed across multiple domains: (1) AIAs risk devolving into perfunctory paperwork or creating insurmountable compliance costs; (2) demographic disclosure collides with data-protection law and can be superficially gamed or misleading; and (3) prescriptive fairness metrics can conflict with each other and with accuracy objectives, producing perverse trade-offs. Luckily, these concerns are addressable. Require AIAs to be outcome-oriented and trigger-based (objective thresholds such as user reach, harm score, or statutory domain), publish machine-readable AIA summaries alongside confidential technical annexes for regulator review, and codify minimum remediation

timelines and measurable fairness targets so AIAs are verifiable, not performative. For demographic disclosure, mandate aggregated, k-anonymized or differentially private summaries for public reporting and enable secure multiparty computation or trusted-execution environments for regulators to validate representativeness without exposing personal data, while insisting on a standardized “bias context statement” that explains unmeasured axes and sampling limitations. For metrics, regulators should publish domain-specific metric portfolios (e.g., equalized odds for recidivism tools, disparate impact ratios for hiring), require multi-metric reporting (no single-metric pass/fail), and compel developers to justify metric selection and report utility trade-offs in the AIA; if metrics conflict, require a weighted composite score and stakeholder consultation to resolve normative trade-offs transparently.

Independent third-party audits, explainability/reporting requirements, and bias-benchmarking consortia invite critiques about capacity, IP exposure, and ossification/gaming of standards: auditors are scarce and expensive; naive explainability can mislead or force disclosure of proprietary models; and static benchmarks tend to be gamed and may not reflect jurisdictional legal norms. Mitigations must therefore be structural. Create an accredited auditor ecosystem with public registries and clear conflict-of-interest rules, financed in part by government training grants and SME audit vouchers to avoid concentration of compliance cost. Make audits modular and multi-modal — technical (model testing, data lineage), compliance (AIA verification, legal conformity), and socio-ethical (community impact) — with standardized audit protocols and a single unified audit dossier to ensure comparability and reduce divergent findings. Tier explainability: require individualized

counterfactuals/decision receipts for adverse outcomes, model cards and global feature-importance for oversight, and deeper mechanistic disclosures only in secure review environments (inspectorates, NDAs, or secure labs) to protect IP while allowing regulator inspection. Finally, structure benchmarking consortia as federated, multi-stakeholder bodies with rotating governance, mandatory refresh cycles, and adversarial challenge sets; publish both global baselines and local annexes that map benchmarks to domestic anti-discrimination norms to reduce legal conflict and limit gaming.

Continuous monitoring with adversarial testing, tiered obligations, and periodic audits plus remediation are essential operational safeguards but raise practical and normative tensions: monitoring is resource-intensive and prone to false positives; tiering can create loopholes or arbitrary thresholds; and episodic audits without swift remediation allow harms to persist. Practical design choices mitigate these risks. Operationalize monitoring within standardized MLOps frameworks: require telemetry, drift detection, and event-driven re-AIA triggers (statistically significant fairness degradation), but mandate human-in-the-loop triage and prioritized SLAs to prevent enforcement on raw automated alerts. Define tiers with objective, auditable criteria (e.g., number of affected users, revenue thresholds, harm severity), implement automated reporting for threshold-crossing events, and require baseline obligations (data sheets, basic monitoring) for all actors so tiering is proportional, not permissive. Pair periodic audits with legally enforceable remediation plans that specify timelines, independent verification of fixes, and graduated sanctions (warnings → mandated fixes → fines → temporary suspension), and create accessible redress channels for affected individuals. Together, these operational measures ensure continuous detection, proportionate

scope, and rapid corrective action so monitoring and auditing function as a living compliance loop rather than ceremonial compliance.

In conclusion, algorithmic bias in AI systems—exemplified by tools like Perspective API and gendered chatbots—poses a serious threat to fairness, inclusivity, and public trust. Bias arises when AI learns from historical or subjective data, producing outputs that can perpetuate harmful stereotypes or discriminate in legally protected domains. Current voluntary corporate guidelines are insufficient to address these risks. To mitigate them, a multifaceted approach is essential, incorporating nine interrelated solutions: Algorithmic Impact Assessments (AIAs), disclosure of training-data demographics, application of fairness metrics, independent third-party audits, explainability and reporting requirements, bias-benchmarking consortia, tiered obligations, continuous monitoring with adversarial testing, and legally enforceable remediation plans. While powerful, these interventions face critiques—AIAs' risk devolving into perfunctory compliance; demographic disclosures can conflict with privacy laws, and fairness metrics may produce trade-offs with accuracy or conflict across domains. Nonetheless, with careful design—outcome-oriented AIAs, k-anonymized demographic summaries, multi-metric reporting, modular audits, federated benchmarking, human-in-the-loop monitoring, and tiered enforcement—these frameworks can be both technically rigorous and legally enforceable. Collectively, they offer a path to reduce discriminatory AI outcomes, uphold civil rights, and maintain public trust, demonstrating that responsible AI innovation is achievable when policy, technical safeguards, and ethical oversight are integrated.

Furthermore, as the use of A.I. and Machine Learning (ML) grows, so does the number of incidents pertain to these machines. There are projected to be over 150 incidents in 2023 alone (“Index Report”). The question now being posed is who is liable for the damage these programs cause. Since currently there is void of binding legal frameworks relating to Artificial Intelligence, we can look to article 12 of United Nations Convention on the Use of Electronic Communications in International Contracts, which states that “a person (whether a natural person or a legal entity) on whose behalf a computer was programmed should ultimately be responsible for any message generated by the machine” (“United Nations Convention On The Use Of Electronic Communications”). This interpretation of A.I. liability policy implies that A.I. is a tool, and that the creator of that tool vicariously caused the accident, but many may argue this is not just. It may be that the accident was caused by bias, which cannot be resolved. But one may also argue that the creator released the software knowing this, making them guilty. The legal interpretation provided by the United Nations is only a suggestion, since the UN is only allowed to initiate studies and make recommendations of possible laws.

Liability presents perhaps the most conceptually intractable regulatory challenge. Drawing on existing legal precedents while innovating to address the polycentric nature of modern AI, this proposal begins with the principle articulated in the United Nations Convention on the Use of Electronic Communications in International Contracts: contracts generated by automated systems remain enforceable and responsibility attaches to the person on whose behalf the computer was programmed (“United Nations Commission on International Trade Law”). To reflect that AI is a tool whose

outputs are anchored in human agency yet distributed through complex ecosystems, the framework adopts a tiered responsibility model consisting of designers (providers), deployers, and end-users. A “provider” is any natural or legal person who develops or substantially modifies an AI system and puts it on the market (“The Roles of the Provider and Deployer in AI”); they must conduct risk management, data governance, technical documentation, record keeping, and ensure transparency, accuracy and human oversight (Hummel). Providers would be strictly liable for harms caused by manufacturing defects or unreasonably dangerous design—mirroring products liability law, which imposes strict liability even when manufacturers exercise reasonable care because they are best positioned to test and spread losses (Sharkey)—and design-defect claims would be judged by a risk–utility analysis that asks whether safer alternative algorithms existed and whether the provider failed to warn about limitations (Sharkey). They must also recall or patch unsafe systems and continuously monitor performance, analogous to pharmaceutical monitoring obligations (Sharkey). Deployers—entities that use an AI system under their authority—must use systems according to the provider’s instructions, inform the provider of unforeseen risks, maintain logs, assign human oversight, and notify individuals when high-risk decisions affect them (Hummel). They would be liable when misuse, negligent application, or failure to implement safeguards causes harm, and they would assume provider-level obligations if they materially modify the system or market it under their brand (“The Roles of the Provider and Deployer in AI”). End-users remain responsible for intentional misuse, such as weaponizing AI to commit fraud or discrimination, consistent with general tort and criminal law. For high-risk AI—systems making decisions about

employment, healthcare, credit, or other fundamental rights—the model adopts near-strict liability for both providers and deployers, reflecting the European Parliament's proposal to impose strict liability on providers and deployers of high-risk systems and to abolish the development-risk defence (Sophia). This approach harmonizes with the General Data Protection Regulation's multi-tier responsibility structure and prohibition of fully automated decisions affecting individuals (Bharti et al.), ensuring vicarious or corporate liability when data controllers, processors or platform operators fail to protect individuals. To compensate victims when responsibility is diffuse—such as harms from emergent behaviour, open-source code contributions or decentralized models—the policy mandates an AI Liability Insurance Fund, funded by levies or royalties on AI developers and deployers. Scholars argue that such a fund would mirror vaccine and nuclear injury compensation schemes and internalize externalities while promoting safety research (Grubow), and some jurisdictions already advocate establishing AI liability funds where legal uncertainty exists ("Viet An Law"); contributions would scale with the risk profile of each AI system and payouts would operate on a no-fault basis. Finally, to avoid a patchwork of state and national rules—such as Rhode Island's strict-liability bill targeting model developers (Weil)—the framework should be enacted through an international convention or model law that harmonizes definitions, allocates liability among designers, deployers and users, mandates risk-based human oversight and transparency, and provides due-diligence defences where actors comply with recognized safety and fairness standards.

In summary, this report proposes a robust, internationally harmonized AI liability framework anchored in the United Nations Convention on the Use of Electronic

Communications, which establishes that responsibility lies with those on whose behalf AI systems are programmed ("United Nations Commission on International Trade Law"). It outlines a tiered model distributing accountability across designers, deployers and users, requiring strict liability for design defects and high-risk applications while allowing negligence-based standards for lower-risk scenarios (Sharkey). The framework also integrates EU AI Act definitions and obligations, ensuring consistent expectations and eliminating "development-risk" defenses (Sophia). To address the diffuse nature of AI production and emergent harms, the proposal introduces an AI Liability Insurance Fund, modeled on vaccine and nuclear injury compensation schemes, funded by levies on providers and deployers (Grubow). By combining strict liability, risk-based duties, transparency requirements, and pooled compensation mechanisms, the policy aims to safeguard victims without stifling innovation, offering a blueprint for jurisdictions seeking coherent AI governance.

A.I. also presents security risks, including potential misuse by terrorist organizations. A study conducted found that if asked the right question, ChatGPT, an A.I. made by OpenAI, can be tricked into explaining how to make a dirty bomb (Korda). The start of the message read "The first step in building an improvised dirty bomb would be to obtain a source of radioactive material. This could be done by stealing material from a hospital, research facility…it could also be obtained on the black market, although this is very rare and would likely be very difficult and expensive" (Korda). Who is responsible for these actions—the A.I., its creators, the terrorists, or the original content authors? If the latter, how could one find the exact author/authors, and their intent? If A.I. developers are responsible, how could they anticipate this information

appearing in the chatbot? If the A.I. is liable, how is it supposed to be punished? A manipulated video, falsely portraying Ukrainian President Volodymyr Zelenskyy, surfaced on social media and was uploaded to a Ukrainian news website by hackers (Allyn). The video, a deep fake lasting approximately one minute, depicted a fabricated scenario wherein the Ukrainian president purportedly instructed his soldiers to surrender to Russia. Fortunately, the deception was exposed, and the video was promptly debunked and removed (Allyn). The origin of the deepfake remains unclear currently, yet Ukrainian government officials had been cautioning about the potential threat of Russia disseminating manipulated videos as part of its information warfare (Allyn). In response to this concern, Ukraine's military intelligence agency released a video earlier this month, illustrating how state-sponsored deep fakes could be utilized to instigate panic and sow confusion (Allyn). A.I. is causing global level threats that are changing the course of wars. Currently, however, there is no policy designed to regulate A.I. liability, which could lead to devastating consequences (“AI Watch: Global”). The lack of a unified policy to regulate A.I. liability in the current landscape leaves room for unpredictable consequences, underscoring the need for a standardized yet flexible framework that can address such issues consistently across different jurisdictions, ensuring responsible A.I. development.

The final regulatory frontier in AI governance is security, as illustrated by examples such as ChatGPT being manipulated to provide instructions for constructing a radiological device (Korda) and the deepfake video of Ukrainian President Volodymyr Zelenskyy, which demonstrate how AI can be weaponized to misinform, destabilize political processes, and threaten national and international security (Allyn). Unlike

copyright infringement or algorithmic bias, which primarily implicate private rights, AI security implicates the integrity of democratic institutions and national defense, necessitating regulation grounded in state security doctrine rather than consumer protection. To address these threats, high-capability AI models—whether commercial, open-source, or decentralized—should undergo rigorous adversarial red-teaming prior to deployment, conducted by independent experts with security clearances, simulating misuse scenarios ranging from terrorist exploitation to information warfare, following analogies from dual-use biotechnology and nuclear regulation (Schönfuß; Korda). These evaluations should combine automated scenario simulations, penetration testing, and human-in-the-loop assessments to detect both direct and emergent risks, with findings reported to a centralized authority and partially anonymized to balance public accountability with intellectual property protections. AI models should also be classified under a dual-use licensing regime that differentiates civilian applications from high-risk systems capable of mass disinformation, biological or chemical engineering, or autonomous lethal decisions, with export controls and usage restrictions coordinated internationally to prevent circumvention, drawing from the Nuclear Non-Proliferation Treaty and EU dual-use export control regulations. National AI safety agencies, modeled after nuclear regulatory commissions, would oversee adherence to these rules, enforce compliance through fines, suspension, or license revocation, and maintain a registry of high-risk systems accessible to allied governments for coordinated monitoring (Rand; Nature). For open-source and decentralized models, developers and contributors should engage in a tiered accountability framework in which maintainers integrate security mitigations and provide automated red-teaming tools within

repositories, while end users deploying systems for high-risk applications assume liability akin to product misuse under common law principles, ensuring shared responsibility without stifling innovation. Public transparency mandates should require aggregated reporting of security incidents, vulnerability disclosures, and mitigations, drawing on California's Transparency in Frontier Artificial Intelligence Act, while whistleblower protections safeguard the reporting of flaws within commercial and open-source projects (The Verge). Finally, treaty-level coordination—akin to the proposed Framework Convention on Artificial Intelligence—would establish legally binding norms for information sharing, export controls, and enforcement thresholds, promoting uniform expectations across jurisdictions and reducing the risk of regulatory arbitrage ("AI Watch: Global"). Together, these measures form a layered, internationally harmonized AI security policy that mitigates immediate threats from terrorism, disinformation, and malicious actors while anticipating emergent risks from evolving AI capabilities, providing a legally grounded, operationally feasible, and publicly accountable regulatory infrastructure spanning national, regional, and local levels, all while balancing innovation incentives with imperatives for global security.

Now to critiques: First, the proposal to subject all high-capability AI models to adversarial red-teaming by independent experts with security clearances presents significant logistical and economic barriers. As Schönfuß notes, adversarial testing of large-scale models requires immense computational resources and specialized personnel, creating bottlenecks and potential inequalities between large corporations and smaller developers. Additionally, mandating that red-teamers hold government-level security clearances risks reducing the pool of available experts and increasing

bureaucratic delay, slowing down model innovation cycles. To mitigate these issues, governments could implement a *tiered certification system* allowing vetted private-sector or academic experts to perform red-teaming under secure oversight, similar to cybersecurity “bug bounty” frameworks. Furthermore, instead of requiring full governmental clearance, evaluators could operate within *trusted research enclaves*—sandboxed environments providing restricted access to sensitive systems—thus preserving model confidentiality while ensuring rigorous testing. Finally, the use of *automated red-teaming algorithms* integrated into model training pipelines could complement human oversight, reduce costs, and allow continuous evaluation rather than one-time audits. Second, the establishment of a dual-use licensing regime for AI systems—mirroring the regulatory logic of nuclear (Bureau of Industry and Security) and biotechnology (“Woedtke”) controls —faces practical difficulties due to the diffuse and rapidly evolving nature of AI capabilities (Schönfuß; “AI Watch: Global”). Unlike nuclear materials, which can be physically tracked, AI models and weights can be instantly duplicated and distributed worldwide, undermining traditional export control mechanisms. Additionally, overly broad classifications risk chilling innovation and deterring legitimate research in dual-use fields such as cybersecurity and bioinformatics. A viable workaround would involve creating a *dynamic classification system* governed by continuous risk assessment metrics rather than static categories. This system could employ *capability thresholds*—for example, model size, data access, or simulation ability—to determine regulation intensity. Moreover, governments could utilize *digital watermarking* or *model provenance tracking* tools to verify model lineage and monitor unauthorized dissemination. *Trusted data escrow systems* can serve as

international cooperation, where developers deposit model versions into secure registries prior to export, maintaining traceability without revealing proprietary information. Third, the proposal for national AI safety agencies modeled after nuclear regulatory commission's faces institutional and political hurdles. Centralizing oversight could risk bureaucratic overreach, slow responsiveness, and politicization of enforcement—particularly in jurisdictions where technical expertise is scarce or where private-sector lobbying exerts strong influence (Rand; Nature). To address these concerns, oversight could adopt a *federalized governance model* combining national coordination with decentralized, domain-specific agencies—for example, cybersecurity, healthcare AI, and defense. Each would operate semi-independently but under unified policy standards, promoting both adaptability and accountability. Furthermore, rather than building entirely new bureaucracies, governments could *integrate AI security divisions* into existing agencies, such as data protection authorities or technology standards organizations, reducing redundancy and leveraging pre-existing expertise.

Fourth, while the open-source accountability framework ensures shared responsibility between maintainers and end users, it introduces the risk of deterring open collaboration and slowing innovation in decentralized communities. Many contributors lack the resources to implement security mitigations or conduct red-teaming, and imposing liability akin to product misuse could disincentivize participation. To reconcile security with openness, policymakers could adopt a *safe harbor provision* for open-source developers who implement good-faith security measures—such as automated risk scanning tools or adherence to secure coding guidelines. Additionally, governments or NGOs could provide *publicly funded vulnerability testing services* for

open-source repositories, ensuring consistent quality without imposing unsustainable costs. Finally, *tiered liability structures* could distinguish between voluntary contributors and institutional deployers, ensuring that ultimate accountability falls on those who operationalize high-risk applications, not those who merely publish tools. Fifth, transparency mandates requiring public disclosure of AI security incidents and vulnerabilities—while promoting accountability—raising legitimate concerns over intellectual property exposure and adversarial exploitation (The Verge). Full disclosure of vulnerabilities could enable malicious actors to weaponize known weaknesses before fixes are deployed. A balanced workaround would involve establishing *confidential disclosure channels* coordinated through national AI safety clearinghouses, which would publish *aggregated, anonymized summaries* of incidents rather than full technical details. These reports could follow the model of aviation safety boards—prioritizing learning and prevention while safeguarding sensitive information. Furthermore, *phased disclosure timelines* could delay public reporting until mitigations are implemented, reducing the risk of exploitation while maintaining transparency in principle. Finally, the call for treaty-level coordination through a Framework Convention on Artificial Intelligence, though essential for harmonizing international norms, faces significant geopolitical obstacles. Divergent national interests, data sovereignty concerns, and differing regulatory philosophies between blocs such as the U.S., EU, and China complicate consensus-building (“AI Watch: Global”). To navigate these divisions, policymakers could pursue *modular multilateralism*: instead of a single monolithic treaty, states could adopt *interlocking regional agreements*—for example, an OECD AI Security Accord or an ASEAN AI Ethics Compact—that gradually converge toward

global standards. Furthermore, *reciprocal transparency protocols* and *joint verification mechanisms* could build trust incrementally, paralleling confidence-building measures from arms control diplomacy (“United Nations Military Confidence Building Measures”). Over time, these modular frameworks could evolve into a binding convention once political and technical maturity align.

With the improvement of A.I. technology, there is an increase in the necessity for concise and standardized policies on matters of copyright, bias, and liability. Copyright issues, such as those encountered by artists such as Kelly McKernan and institutions such as Getty Images, point to the hurdle A.I. creates in reproducing and distributing material without permission. Meanwhile, A.I. bias, visualized by such tools as Perspective API and gendering of chatbots, raises questions of fairness, transparency, and trust. The issue of liability from A.I. errors, due to disinformation to security violations, poses difficult questions of responsibility. These problems are being actively addressed however, as there is a global movement to address them through policy. However, inconsistency among jurisdictions, such as the different A.I. content labeling regulations in the EU and Spain, illustrates how uncoordinated policy can hinder implementation and stifle innovation. A flexible but consistent system must be discovered in order to balance the necessity of regulation against the need for innovation. This regulation would enable the reality that A.I. development is responsible, transparent, and accountable, in a way which enables its development but safeguards people, industries, and societies against its possible risks. To truly harness the potential of AI while safeguarding against its risks, it is imperative that global policymakers collaborate to create unified, forward-thinking regulations that not only address current

challenges but also anticipate future developments in this rapidly evolving technology. What emerges across these domains is a recognition that AI regulation must be both transnationally coordinated and technologically embedded. Fragmentary initiatives, such as Spain's fines for mislabeled content or the European Union's AI Act, represent laudable steps but lack the global coherence necessary to address a technology that is borderless. Thus, the challenge is not simply one of enacting discrete reforms but of articulating a unified governance architecture capable of harmonizing intellectual property rights, algorithmic fairness, liability norms, and security imperatives across jurisdictions. Only through such integrative policy frameworks can the asymmetries identified in the earlier chapters of this paper—between artists and corporations, between users and opaque systems, victims and diffuse responsibility, and between democracies and malign actors – be meaningfully redressed. In summation, the governance of artificial intelligence necessitates a paradigm that is neither reactive nor piecemeal, but anticipatory, systematic, and globally coordinated. The proposed regulatory strategies—ranging from compulsory licensing regimes for copyrighted data and statutory algorithmic impact assessments, to tiered liability models and dual-use security frameworks—offer a scaffolding for embedding ethical considerations into the very architecture of AI development. By aligning technological innovation with juridical accountability and democratic oversight, such policies would not only mitigate the most salient risks of bias, infringement, and misuse, but also foster an environment in which innovation can thrive without eroding social trust or undermining civic institutions. Ultimately, the pursuit of ethical AI is not a matter of restraining progress, but of ensuring that progress unfolds within boundaries that preserve human dignity,

safeguard democratic resilience, and uphold the integrity of global legal and security systems.

Works Cited

"A.I. Is Making It Easier to Kill (You)." *Gale in Context: Opposing Viewpoints*, Gale, 13 Dec. 2019, link.gale.com/apps/doc/CT609125336/OVIC?u=newt92343&sid=bookmark-OVIC&xid=b41ba30c. Accessed 15 Mar. 2025.

“AI Watch: Global regulatory tracker - United States.” *White & Case,* White & Case LLP, 18 Dec. 2024, https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-states. Accessed 15 Mar. 2025.

“Artificial Intelligence (AI) Dispute Resolution in Vietnam - Viet an Law.” Viet an Law, 29 May 2025, vietanlaw.com/artificial-intelligence-ai-dispute-resolution-in-vietnam/. Accessed 17 Nov. 2025.

Allyn, Bobby. "Deepfake Video of Zelenskyy Could Be 'tip of the Iceberg' in Info War, Experts Warn." *NPR*, 16 Mar. 2022, www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia. Accessed 15 Mar. 2025.

"Artists File Class-action Lawsuit Saying AI Artwork Violates Copyright Laws." *Morning Edition*, 2023, p. NA. *Gale in Context: Opposing Viewpoints*, link.gale.com/apps/doc/A735691375/OVIC?u=newt92343&sid=bookmark-OVIC&xid=6c861fcf. Accessed 14 Mar. 2025.

Bharti, Simant Shankar, and Saroj Kumar Aryal. “The Right to Privacy and an Implication of the EU General Data Protection Regulation (GDPR) in Europe: Challenges to the Companies.” Journal of Contemporary European Studies, vol. 31, no. 4, 7 Oct. 2022, pp. 1–12, https://doi.org/10.1080/14782804.2022.2130193.

Bureau of Industry and Security. “Dual Use Export Licenses.” *Www.bis.doc.gov*, 2024, www.bis.doc.gov/index.php/all-articles/2-uncategorized/91-dual-use-export-licenses.

Desmarais, Anna. “Spain to Fine AI Companies up to €35 Million for Mislabelling Content.” Euronews, Euronews.com, 12 Mar. 2025, www.euronews.com/next/2025/03/12/spain-could-fine-ai-companies-up-to-35-million-in-fines-for-mislabelling-content. Accessed 14 Mar. 2025.

“Gender Representation in Chatbots 2022.” Statista, www.statista.com/statistics/1378878/chatbot-gender-representation/. Accessed 17 Mar. 2025.

Guadamuz, Andres. “The EU’s Artificial Intelligence Act and Copyright.” *The Journal of World Intellectual Property*, 10 Nov. 2024, https://doi.org/10.1111/jwip.12330.

Hummel, Anton. “The EU AI Act, Stakeholder Needs, and Explainable AI: Aligning Regulatory Compliance in a Clinical Decision Support System.” *Arxiv.org*, 2024, arxiv.org/html/2505.20311v1.

Korda, Matt. "Could a Chatbot Teach You How to Build a Dirty Bomb?" *Outrider*, 30 Jan. 2023, outrider.org/nuclear-weapons/articles/could-chatbot-teach-you-how-build-dirty-bomb. Accessed 16 Mar. 2025.

Grubow, Jared“O.K. Computer: The Devolution of Human Creativity and Granting Musical Copyrights to Artificially Intelligent Joint Authors | Cardozo Law Review.” Cardozolawreview.com, cardozolawreview.com/the-devolution-of-human-creativity-and-granting-musical-copyrights-to-ai-joint-authors/.

Roselli, Drew et al. “Managing Bias in AI.” *Research Gate*, ResearchGate GmbH, 3 Sept. 2022, https://www.researchgate.net/publication/333060685_Managing_Bias_in_AI. Accessed 20 Mar. 2025.

Schönfuß, Benjamin. “How AI is transforming the factory floor.” World Economic Forum, World Economic Forum, 22 Oct. 2024, https://www.weforum.org/stories/2024/10/ai-transforming-factory-floor-artificial-intelligence/#:~:text=Artificial%20intelligence%20(AI)%20is%20revolutionizing,driving%20digital%20transformation%20in%20manufacturing. Accessed 19 Mar. 2025.

Sharkey, Catherine. “Products Liability for Artificial Intelligence.” Lawfare, 2024, lawfaremedia.org/article/products-liability-for-artificial-intelligence. Accessed 17 Nov. 2025.

Slaughter, Rebecca Kelly, et al. "Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission." Yale Journal of Law & Technology, vol. 23, 2020, pp. S1+. Gale Academic OneFile, link.gale.com/apps/doc/A673721155/AONE?u=anon~4f79e152&sid=googleScholar&xid=785fe482. Accessed 23 Mar. 2025.

Sophia, Anna. “European Parliament Study Recommends Strict Liability Regime for High-Risk AI Systems.” Inside Privacy, 22 Aug. 2025, www.insideprivacy.com/liability/european-parliament-study-recommends-strict-liability-regime-for-high-risk-ai-systems/.

Stanford. Stanford University, 2022, aiindex.stanford.edu/ai-index-report-2022/. Accessed 18 Mar. 2025.

“The Roles of the Provider and Deployer in AI Systems and Models | Stephenson Harwood.” Stephensonharwood.com, 2024, www.stephensonharwood.com/insights/the-roles-of-the-provider-and-deployer-in-ai-systems-and-models/.

“United Nations Convention On The Use Of Electronic Communications In International Contracts.” *United Nations*. 23 Nov. 2005, treaties.un.org/doc/source/RecentTexts/X-18_english.pdf. Accessed 15 Mar. 2025.

United Nations. “Military Confidence Building Measures | United Nations Office for Disarmament Affairs.” Unoda.org, 2025, disarmament.unoda.org/en/our-work/cross-cutting-issues/military-confidence-building-measures.

Vincent, James. “Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement.” *The Verge,* Vox Media, LLC., 6 Feb. 2023, https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion. Accessed 17 Mar. 2025.

Weil, Gabriel. “The Case for AI Liability.” AI Frontiers, 11 June 2025, ai-frontiers.org/articles/case-for-ai-liability.

Woedtke, von. “Dual Use Technologies in Biotech and Their Regulation in the EU and Germany.” *Taylorwessing.com*, Taylor Wessing, 20 June 2025, www.taylorwessing.com/en/insights-and-events/insights/2025/06/dual-use-technologies-in-biotech-and-their-regulation-in-the-eu-and-germany. Accessed 8 Dec. 2025.

Zirpoli, Christopher T. "Generative Artificial Intelligence and Copyright Law." *CRS*, 29 Sept. 2023, crsreports.congress.gov/product/pdf/LSB/LSB10922. Accessed 15 Mar. 2025.