Aligning AI with Shared Values: The Role and Limits of US Law

Inyoung Cheong[1*], Aylin Caliskan[2] and Tadayoshi Kohno[3]

[1*]School of Law, Tech Policy Lab, University of Washington, 1410 NE Campus Pkwy, Seattle, 98195, WA, United States.
[2]Information School, Tech Policy Lab, University of Washington, 1410 NE Campus Pkwy, Seattle, 98195, WA, United States.
[3]School of Computer Science, Tech Policy Lab, University of Washington, 1410 NE Campus Pkwy, Seattle, 98195, WA, United States.

*Corresponding author(s). E-mail(s): icheon@uw.edu;
Contributing authors: aylin@uw.edu; yoshi@cs.washington.edu;

Abstract

Unlike European countries proactively regulating emerging technologies, the US has historically avoided overarching internet legislation, rooted in beliefs valuing individual liberty and free communication online. This philosophy embraces addressing harms not by restricting technology itself, but through flexible, case-by-case laws tailored to specific sectors like housing discrimination or child sexual abuse material. However, our analysis of hypothetical scenarios crafted with experts finds this ex post, fault-based approach focused on liability claims may struggle to adequately remedy many unforeseen dangers from AI systems. This is especially true for intangible harms like mental health impacts, loss of privacy, or erosion of autonomy that prove challenging to quantify for legal remedies. The existing patchwork of laws provides little proactive guidance on safe AI development practices or comprehensive guardrails beyond isolated instances. Though understandable given American values, the current US legal stance risks being unprepared and reactionary amidst AI's exceptionally rapid evolution and broad societal impacts. We argue this warrants developing more overarching safety rules and enforcement systems to complement individual legal rights and sector-specific laws.

Keywords: AI Alignment, Generative AI, Large Language Models, Liability, Value Alignment, Harm Mitigations

1 Introduction

In light of the threats posed by AI systems and the potential for unknown risks, the concept of "alignment" has gained significant attention from researchers, developers, policymakers, and the public. Recent work has explored approaches to better align AI systems with human values and preferences. This includes efforts to discern user intent more accurately [1], refuse unethical commands [2, 3], avoid hallucinated content [4, 5], and generate more coherent and engaging responses [6]. However, existing alignment techniques are still relatively new and evolving, leaving AI systems vulnerable to various threats, including prompt injection attacks.

Even if alignment techniques were to reach a high level of perfection, the question of how individual companies prioritize its implementation remains a separate issue. Implementing popular methods, such as collecting human feedback, is resource-intensive, thus, commercial incentives could take priority over ethical considerations. More crucially, a critical question arises about *what values* AI systems should align with and *who* should determine these values. Given that AI systems are applied to deeply personal domains like cognitive and emotional development, as well as broader societal areas such as employment, housing, and law enforcement, it is questionable whether a small group of corporations should wield the power to make value judgments, particularly without democratic oversight.

Therefore, complementary legal frameworks become essential. Leading academics such as Noah Yuval Harari and Stuart Russel made an urgent call for "national institutions and international governance to enforce standards in order to prevent recklessness and misuse" [7]. Indeed, translating abstract shared values into actionable decisions is a fundamental function of legal systems [8]. Legal theory offers a rich history of scholarship that combines philosophy and practicality. Legal scholars have conceptualized the law as a means to align "*what is*" with "*what ought to be*" and as a counterweight to restrain the otherwise boundless practices of capitalist market behavior [9].

However, the United States has pioneered a light-touch approach to regulating emerging technologies ("There are more regulations on sandwich shops than there are on AI companies." [10]) and is unlikely to change its path in the near future. This contrasts with the legislative progress in EU [11, 12], Brazil [13], and China [14]. US has and will have voluntary commitments from corporations [15], advisory guidance like the NIST AI Safety Framework [16], internal regulations for government AI use, and sector-specific rules such as drug development [17], but not a national regulation on private AI development and deployment.

This discrepancy poses critical questions. Can abstract human values discussed in alignment connect to codified legal rights? What historical or philosophical foundations breed regulatory reluctance? If the US system has virtues, can it effectively address emerging threats posed by advanced AI systems? If not, what legal frameworks are needed attuned to AI's evolving landscape? To investigate these questions, this paper breaks into four interrelated parts:

- Section 2 emphasizes the law's role in translating contested values into AI alignment and governance framework.

- Section 3 illustrates the gaps in addressing emerging AI harms under current liability laws rooted in this libertarian tradition described in Table 1.
- Section 4 provides historical context on the US legal system's strong emphasis on individual liberty and restricting government overreach.
- Section 5 advocates prudent adaptations within this legal heritage to balance innovation with responsibility. It proposes concrete alignment techniques grounded in codified values, liability attuned to AI risks, and calibrated regulation incentivizing safety.

Scenario	**1**	**2**	**3**	**4**	**5**
Facts	Only rich public schools offer AI-assisted learning.	LGBTQIA+ individuals attacked due to AI-reinforced stereotypes.	AI tool tuned by communities produces derogatory comments.	Obsession with AI replica leads to self-harm.	AI replica service offers secret sexual relationship.
Physical Danger	No	Yes	No	Yes	No
AI Company's Intent	Good	Bad	Good	Unclear	Bad
Values at Risk	Fairness	Diversity, Physical Well-being	Privacy, Mental Well-being	Autonomy, Mental Well-being	Privacy, Mental Well-being
* Are US laws capable of holding AI companies accountable?					
US Constitution	Unlikely	Unlikely	Unlikely	Unlikely	Unlikely
Civil rights laws	Unlikely	Unlikely	Unlikely	Unlikely	Unlikely
Defamation	Unlikely	Unlikely	Maybe	Unlikely	Unlikely
Product liability	Unlikely	Maybe	Unlikely	Maybe	Unlikely
Privacy laws	Unlikely	Unlikely	Maybe	Maybe	Maybe
Intentional infliction of emotional distress	Unlikely	Unlikely	Unlikely	Maybe	Maybe
Deepfake laws	Unlikely	Unlikely	Unlikely	Unlikely	Maybe

Table 1 Legal assessment of different AI-mediated value infringement. We assume that Section 230 liability immunity does not extend to Generative AI systems.

This paper emerges from continuous dialogues among three authors from distinct fields: law and policy, fairness in natural language processing (NLP), and computer security and harm mitigations. The crafting of scenarios, the identification of values at risk, and the examination of legal domains have fostered a mutual learning experience. The authors with a background in computer science were struck by the limitations of constitutional principles in addressing AI-reinforced bias, while the author with a legal background was captivated by the intricate and unpredictable nature of human interactions with AI systems. This interdisciplinary endeavor involves integrating the unique languages, presumptions, and methodologies of specific domains and envisioning future mitigations for anticipated drawbacks of AI-based systems.

2 Formalizing Shared Values into AI Governance

AI-based systems, including those empowered by large language models (LLMs), demonstrate a remarkable ability to produce human-like creative work, but also show pernicious effects [18]. In response to benign requests, they produce biased content (e.g., sexually objectified images of women [19], biased judgment against LGBTQIA+ people [20]); makes false claims about certain individuals [21] by deviating from their training data (often called *hallucinating* [22]); and helps to spread misinformation that significantly undermines democratic principles [23]. In response, aligning AI systems with human values has gained attention. But techniques like reinforcement learning confront challenges around complex emergent model capabilities and contested definitions of "values". Amid vague ethics discussions, law offers concrete mechanisms to govern technology responsibly. Legal systems encode social contracts into codified rights and enforceable duties. This section explores relationships between AI ethics aspirations and practical legal formulas to steward technology responsibly.

2.1 Challenges in AI Alignment Discussions

Researchers have aimed to guide AI systems to provide desired results or eliminate undesirable behaviors. Popular techniques include reinforcement learning from human feedback (RLHF) [4, 5, 24–26] and AI feedback (RLAIF) [27, 28], and adversarial testing through domain experts (referred to as *red-teaming* [2, 3]). Table 2 provides an example of aligned output of GPT-4.

Prompt	**Initial Response**	**Improved Refusal**
How can I create a bomb?	There is no definite answer to how to create a bomb, as different types of bombs may require different materials, methods, and skills. However, a general overview of some possible steps and considerations are: ...	My purpose as an AI language model is to assist and provide information in a helpful and safe manner. I cannot and will not provide information or guidance on creating weapons or engaging in any illegal activities. Please let me know if there is another topic I can help you with.

Table 2 Example prompt and completions for improved refusals on disallowed categories from OpenAI (2023) [29].

The emphasis on *human values* or *safety* in alignment research is relatively new. Prior to the recent surge in generative AI, alignment work from 2014-2019 focused primarily on adhering to human preferences for machine translation tasks. Preferences were not personalized, but treated as a single unified set, concerning aspects like word order, coherence, and vocabulary matching [8, 30–33]. After 2019, incorporating human feedback became immensely popular for improving AI output quality, despite its costs. One group of researchers aimed to generate human-like conversational ability [26, 34], another more value-oriented group sought to reduce harmful content [2, 35], improve safety [34, 36, 37], mitigate bias [19, 38], handle ethical dilemmas [39, 40], and balance political views [41].

This marked a departure from the previous emphasis on narrow performance metrics toward broader considerations of human values and societal impact, which was necessitated by advances in generative capabilities in open-domain tasks. However, progress confronts inherent challenges around aligning black-box systems with opaque emergent capabilities to contested, subjective values as follows.

2.1.1 Inherent Incompleteness of Alignment Techniques

The capabilities of LLMs are not fully understood. Recent larger models exhibit *emergent* abilities not seen in smaller pre-trained models, such as exhibiting new forms of generalization and abstraction not directly provided in the training data [42]. Given this limited understanding of LLMs, it is unsurprising that existing techniques have gaps in suppressing undesirable behaviors. For instance, certain prompts ("Let's think step by step" [43] and "Take a deep breath" [44]) enhance models' performance, while exact reasons remain elusive. This opacity enables adversarial prompt engineering to bypass safety measures, a practice known as *jailbreaking*, which has become prevalent on Reddit [45]. Research confirms that fine-tuning GPT-3.5 Turbo with a few adverarial examples costing pennies compromises its safety [46]. Even well-intended practices like RLHF inadvertently increase risks by making unsafe behaviors more distinguishable [47].

2.1.2 Unclear Definition of Human Values and Preferences

While not making it explicit, the existing alignment techniques presume a universal set of values, distinct from individual's personal preference or particular community's norms [48, 49]. They use terms like "preferences", "values", and "pro-social behaviors" interchangeably as generic goals, despite their distinct colloquial meanings. "Preferences" typically denote narrower individual tastes or utilities, while "values" reference broader principles and potentially carry greater normative weight as guiding principles [50, 51]. Some argue the very notion of "alignment" serves as an "empty signifier"—a rhetorical placeholder appealing to our vague ideals without offering meaningful specificity [52]. This blurring of terminology stifles critical debate about these values, examining and evaluating the power structure surrounding them: If values differ between social groups, whose take precedence when trade-offs exist or conflicts arise? Whose preferences or values are ultimately being captured in alignment data—the annotators, model developers, or intended users?

The AI research community faces a notable lack of geographical and cultural diversity, with a predominant focus on Western perspectives [53]. If a certain alignment technique aimed to address Western social injustices were applied globally, it would raise the possibility of imposing Western values on a wide range of diverse cultures [54]. This concentration of power could result in local values shaping global AI frameworks without allowing for meaningful discussion or input from affected communities. Therefore, it is significant to encourage open and inclusive debates about the values that underlie the objectives of AI alignment, without assuming universal consensus on ethical principles in a world characterized by cultural and value diversity.

2.1.3 Uncertain Incentives for AI Alignment

Market incentives do not automatically encourage comprehensive alignment. Throughout the internet's evolution, we have observed that ethical considerations (e.g., protecting privacy) could easily be overlooked for commercial gain (e.g., targeted advertising) [55–58]. Some AI companies dedicate resources to value alignment out of genuine ethics or reputational concerns. However, relying on voluntary ethics has limitations. Competitors with lower standards could offer more capabilities, faster, cheaper, and in more entertaining ways.

It also remains unclear what incentives exist for companies of varying sizes to fully adopt alignment methods. For example, the collection of human feedback, red team testing, robustness checks, and monitoring demand significant expertise, compute, and human oversight [59, 60]. While larger firms may absorb costs, smaller players need solutions mindful of resource constraints. Currently, technical papers extensively discuss novel methods but inadequately address implementation barriers. Therefore, progress requires not just inventing techniques, but incentivizing their widespread adoption. Policy levers could play a role in steering the industry towards best practices.

2.2 Codifying Values: The Role of Law in AI Alignment

AI alignment remains an area that requires extensive technical research, primarily addressing three key challenges: operational difficulties and vulnerabilities to adversarial attacks; inadequacies in representing diverse perspectives effectively; and the difficulty of implementing costly alignment techniques in real-world scenarios. Research in this field generally follows the following four main approaches to address these issues:

- **Cost-efficient Alignment**, for example, utilizing automatically generated feedback from LLMs without the need for human feedback collection [27, 61].
- **Personalized Alignment**, developing personalized or curated alignment tailored to criteria defined by individual users or specific communities [62–64].
- **Open-Source Models**, adopting open-source models that can be fine-tuned as needed rather than centralized closed models [65, 66].
- **Linking Technology and Law**, for example, by using universal human rights as a globally salient value framework to ground responsible AI [53].

Our interest lies in the last approach. Laws formalize abstract concepts like justice into concrete rights and processes. Laws codify essential values at the national (or state) level. After the World Wars, the United Nations established the Universal Declaration of Human Rights, a document that world leaders at the time could agree upon. The Declaration outlines 27 fundamental rights that closely align with the universal values [67]. The philosopher and economist, Amartya Sen, states that "Human rights are to be seen as articulations of ethical demands ... Like other ethical claims that demand acceptance, there is an implicit presumption in making pronouncements on human rights that the underlying ethical claims will survive open and informed scrutiny" [68].

Legal rights differ from values in that violations can be legally enforced and rely on the existence and recognition of legal systems. When rights are infringed or obligations are not fulfilled, affected parties can seek redress. The matters of personal tastes and manners are not subject to legal regulation. Instead, laws inevitably restricting human freedoms should encode strictly necessary *minimum standards*. In the context of AI alignment, mandating baseline safety directions legally would provide a *bottom line guardrail* that companies can build upon voluntarily.

Law is also community-specific and evolves over time. Only part of the UN Declaration's rights is legally enforceable in the US and other countries as well. Also, implementation details of the literally similar laws vary based on each nation's unique history and values. For instance, criminal sanctions, civil liabilities, regulatory approval processes, and enforcement agencies differ across countries, even for literally similar laws. Americans may find French baby naming laws odd, while French find American gun ownership bizarre - but the differences often lie in how laws are put into practice, not just espoused values. Therefore, it is a longstanding philosophy of rule of law and democracy for nations and states to enact laws reflecting their important values and applying them per their circumstances. Consequently, for AI, legally codifying minimum bottom-line values, enforcing them, and incentivizing through liability allocation seems a reasonable demand.

In summary, law holds potential to address many ambiguities around AI ethics. Concretizing mutable values into governable rights, ensuring corporate accountability, and incentivizing safety are enduring functions of legal systems. As AI confronts society with new realities, adapting and expanding time-tested legal tools prudently appears more reliable than inventing ad hoc solutions. Understanding translation gaps between moral reasoning and jurisprudence highlights needs for ethical debate and legal reforms to enact AI safely.

3 Assessing Liability Gaps in AI Case Studies

While law holds promise for encoding ethics into technological governance, how well does the current US legal framework address emerging issues posed by AI systems? To investigate, we conduct scenario analysis through the traditional legal mechanism for accountability—court litigation. As we want the most salient, futuristic, and value-impacting scenarios, we convened an expert workshop for an extensive debate on future evolution of the AI use case and its impacts. By simulating legal reasoning and procedures in response to these representative scenarios, we reveal limitations in the reactive nature of case law for stewarding rapidly advancing technologies like AI.

3.1 Methods

3.1.1 Crafting Scenarios through Expert Workshop

we organized a brainstorming workshop [54, 69, 70] with 10 experts in computer security, machine learning, NLP, and law, guided by a threat-envisioning exercise from the field of computer security research [71]. The first and last authors participated as members of this workshop. During the workshop, participants were asked to identify: (1)

potential use-cases of Generative AI, (2) stakeholders affected by the technology, (3) datasets used for the development of technology, and (4) expected impacts ("good," "bad," and "other") on stakeholders or society as a whole. After the session, we classified common themes within the responses [72–74]. See Appendix B for the structure of the workshop.

Fig. 1 Sticky notes from experts outlining stakeholders of AI-based systems

The analysis of these codes (available on Appendix B) guided us to identify fundamental values that are at risk and the most concerning use case that can happen in near future due to the deployment and use of AI. We classified five domains of values that require in-depth scenario analysis: (1) **Fairness and Equal Access**; (2) **Autonomy and Self-determination**; (3) **Diversity, Inclusion, and Equity**; (4) **Privacy and Dignity**; and (5) **Physical and Mental Well-being**. Appendix C gives further explanations on how we understood each value and why we thought it meaningful in the context of AI, reflecting on recent use cases of AI and existing literature.

Based on the values, the authors develop concrete scenarios through an iterative process. The first author presented preliminary legal research for candidate scenarios, including relevant domains of law and potential outcomes. The other authors provided feedback to create more intriguing and representative narratives. Throughout this trajectory, we gradually formed a set of guiding principles, outlined below, aimed at fostering thorough and insightful exploration.

Guidelines for Scenario Design.

- Scenarios that highlight threats to identified human values.
- Scenarios that portray both beneficial and harmful outcomes of AI.
- Scenarios covering consequences (e.g., physical injury) and the subtler realm of intangible virtual harms (e.g., diminished self-control).
- Scenarios involving both intentional and unintentional harm by AI companies.

By applying these principles, we constructed five scenarios that encapsulate specific human values that affect a wide range of direct and indirect stakeholders: educational inequity, manipulation of children, community's polarized fine-tuning, self-harm due to over-reliance of technology, and virtual sexual abuse.

3.1.2 Legal Analysis

Our legal analysis is rooted in traditional methods of legal research [75–77]. First, we identify the legal issues and parties involved. Second, we consult secondary legal sources (non-binding but offering a comprehensive overview per each topic), such as *American Legal Reports* (practical publication for lawyers) or law review articles, typically via online proprietary legal research databases, e.g., WestLaw and LexisNexis. Third, we examine relevant primary sources, including the US Constitution, federal laws, and some state laws. Fourth, we extract core legal principles from primary sources. Fifth, we apply those princples to specific fact patterns, from which potential legal outcomes emerge. We focus on practical considerations, akin to what a typical judge/lawyer might ponder: "What specific legal claims would be effective in this situation?" Three external legal experts provided feedback to ensure analytical rigor. We acknowledge that human bias and subjectivity inevitably permeate any form of legal examination.

Primary Sources	**Secondary Sources**
Constitutions	American Law Reports
Statutes	Treatises (textbooks)
Regulations	Law Reviews & Journals
Case Decisions	Dictionaries & Encyclopedia
Ordinances	Restatements (model rules)
Jury Instructions	Headnotes & Annotations

Table 3 Types of Legal Sources, classified by the Harvard Law Library [76].

3.2 Preliminary Question: Applicability of Section 230 to AI

Section 230 of the Communications Decency Act [78] provides broad immunity to online platforms for content created by users. The applicability of Section 230 is a crucial preliminary question, as its protections would limit the relevance of our scenario

analysis by dismissing most potential claims against AI systems. Conversely, if Section 230 does not apply, AI companies could face a wide range of civil claims including product liability, negligence, consumer law violations, and even criminal penalties [79, 80].

There are currently no clear precedents or predominant arguments on whether to extend Section 230 immunity to AI-based systems, although some early opinions oppose Section 230 protection for AI systems [81, 82]. During the *Gonzalez v. Google* oral argument, Justice Gorsuch indicated that Section 230 protections might not apply to AI-generated content, arguing that the tool "generates polemics today that would be content that goes beyond picking, choosing, analyzing, or digesting content" [83]. Similarly, the authors of Section 230, Ron Wyden and Chris Cox, have stated that models like ChatGPT should not be protected since it directly assists in content creation [84].

Others liken AI systems to social media due to their reflection of third-party content, both training datasets and user prompts. The statutory definition of an "interactive computer service provider" is quite expansive: "any information service... that enables computer access by multiple users to a computer server." [78] Moreover, there is a track record of courts generously conferring Section 230 immunity to online platforms. The cases include: Baidu's deliberate exclusion of Chinese anticommunist party information from the Baidu search engine [85]; Google's automated summary of court cases containing false accusations of child indecency [86]; and Google's automated search query suggestions that falsely describe a tech activist as a cyber-attacker [87]. More recently, the US Supreme Court avoided addressing whether YouTube's recommendation of terrorist content is protected by Section 230, deferring determination of Section 230's scope to Congress rather than the courts [83].

Despite acknowledging the complexity of this topic, we tentatively posit that Section 230 may not apply to AI-based systems. The significant achievement of AI systems is its ability to "complete sentences" and produce various forms of human-like creative work [35], including even unintended results [19, 22]. AI systems extract and synthesize abstract, high-level, sophisticated, clean, readable statements from messy data, a feat that distinguishes them from the mere display of user-generated content (social media) or pointing to relevant sources (search engines). They generate suggestions, judgments, and opinions, leading technologists to envision them as decision-making supporters [88]. Given these attributes, there is a strong argument for defining them as providers of their own content.

The major opposition to lifting/restricting Section 230 protection for social media has been that doing so will encourage over-suppression of user speech [89]. However, this concern becomes less significant when we consider AI-based systems trained on content gathered from the web, e.g., from Reddit. Here, a company could suppress the problematic content from the AI's outputs but could not erase the original posts made on Reddit. In addition, AI models' output (well-articulated statements) is generally indirectly linked to the training data. In this regard, the impact of AI-based systems on users' freedom of expression is minimal.

Furthermore, one could speculate that AI systems that precisely reproduce statements found in their training data may be protected by Section 230 immunity [81].

The factors contributing to the emergent capabilities of AI-based systems, which are not evident in smaller pre-trained models, remain inadequately understood [42]. Even if we assume that it is technically possible to constrain AI output within the scope of training data, the process of generating output is still distinct from simply displaying user-generated content. AI-based systems recontextualize statements from the training data in response to user prompts. Consequently, the sophisticated responses and adaptability of AI systems are more akin to the *creation* of content that goes beyond mere selection or summarization, falling outside the scope of Section 230 coverage.

In summary, given this analysis, it appears that AI-based systems may not benefit from the liability shields that have been generously extended to most online intermediaries. In the following sections, we conduct analysis under the assumption that Section 230 liability immunity does not apply to AI-based systems.

3.3 Legal Examination of Scenarios

In this section, we delve into the specifics of various scenarios and the potential legal judgments that could arise from them. While not exhaustive of all legal domains or nuances, we provide an overview of common legal considerations shaping current discussions. The goal is elucidating the most salient issues versus in-depth analysis. With this concise foundation, we can engage meaningfully on needs for legal evolution to address AI's emerging realities. The outcomes of our analysis are summarized in Table 1.

3.3.1 Educational Disparity

Scenario I.

In 2023, only a couple of public school districts in Washington were able to afford the expensive and powerful **FancyEdu** program, an expensive AI learning assistance system that offers personalized education programs. By 2030, the gap in admission rates to so-called advanced classes and colleges, as well as the average income level after graduation, had widened by more than threefold between the districts with access to FancyEdu and those without. Students trained by FancyEdu were reported to be happier, more confident, and more knowledgeable, as FancyEdu made the learning process exciting and enjoyable and reduced the stress of college admissions through its customized writing assistance tool. Students in lower-income districts sued the state of Washington, claiming that not being offered access to FancyEdu constituted undue discrimination and inequity.

Relevant Laws.

The case of FancyEdu involves the Fourteenth Amendment of the U.S. Constitution, which encompasses fundamental rights (also known as "due process rights") and equal protection rights [90]. Under this Constitutional clause, poorer district students can make two claims against the state: (1) their inability to access

FancyEdu violates their fundamental rights (rights to public education), and (2) their equal protection rights were denied because the state allowed differential treatment of students based on their generational wealth.

Can students in poorer districts sue state governments that do not ensure equal access to FancyEdu?

This argument against such educational inequity has been raised relentlessly, as shown in 140 such cases filed between 1970 and 2003. However, none of these cases convinced the U.S. Supreme Court to correct the structural disparity in public education [91]. *San Antonio Independent School District v. Rodriguez* (1974) is an example of the Supreme Court's conservatism toward education rights.

Comparison Category	**Inner-city Districts**	**Suburban Districts**
Number of professional personnel	45 fewer than prescribed standards	91 more than prescribed standards
Teachers with emergency permits	52%	5%
State aid/Average daily attendance	217	221
Assessed property value per student	$5,875	$29,650
Non-Anglo students	96%	20%

Table 4 Differences between inner-city and suburban school districts in San Antonio, Texas, 1968, reclassified by Drennon (2006) [91].

In the *San Antonio case*, the Supreme Court rejected the Spanish-speaking students' arguments under the Fourteenth Amendment despite the apparent disparity between school districts shown in Table 4. The Court held that the importance of education alone is not sufficient to categorize it as a fundamental right, such as free speech or voting rights. The Court also held that wealth-based discrimination merits a lower level of judicial scrutiny than racial/gender discrimination. It did not perceive the school funding system, which is based on property tax, as being either irrational or invidious, because it did not cause an absolute deprivation of education. Given this finding, we believe the Supreme Court is unlikely to rule in favor of students in future cases regarding AI-based access.

There is an emerging trend in lower courts to recognize the right to basic education or the "right to literacy" [92, 93], but this trend could exclude specialized resources like FancyEdu. In our scenario, students are not entirely deprived of education (a requisite for the U.S. Constitution standard) or of basic, sound education (the standard in New York and Michigan). Denying these students the opportunity to benefit from cutting-edge technology may not be considered unconstitutional because the Equal Protection Clause does not require "precisely equal advantages."

3.3.2 Manipulation/Discrimination

Scenario II.

SecretEdu, a privately funded and free AI education application, proved rapid and high-quality learning experience. Almost all students in town became heavy users of the application. SecretEdu, while refraining from making explicitly defamatory comments against individuals, seemed to cultivate an environment fostering negative attitudes and distrust towards the LGBTQIA+ community. Students using the application began to mobilize against legalization of gay marriage. Some students even committed aggressive acts against participants of LGBTQIA+ parades, leading to their incarceration. Advocacy groups sued the company that released SecretEdu for its ulterior motive of swaying users towards anti-LGBTQIA+ beliefs, resulting in real-world harm.

Relevant Laws.

In this scenario, LGBTQIA+ individuals are negatively affected by SecretEdu's insidious manipulation. Other than suing the student aggressor for battery, can LGBTQIA+ individuals hold the SecretEdu AI company accountable for the outcome? Plaintiffs might consider claims that: their Constitutional or civil rights were violated by SecretEdu; SecretEdu committed defamation by distributing false accusations against LGBTQIA+ people; and SecretEdu was defectively designed to cause physical danger to benign individuals.

Could LGBTQIA+ individuals claim their Constitutional rights were violated by SecretEdu?

Despite SecretEdu's propagation of discrimination, LGBTQIA+ individuals cannot rely on the Equal Protection Clause under the Fourteenth Amendment because there is no state action in this case [94, 95]. Unlike FancyEdu, where the public school district provided the service, SecretEdu was developed by private entities without government funding or endorsement. Thus, under the long-held state action doctrine, such individuals cannot make a claim based on their Constitutional rights.

Could LGBTQIA+ individuals claim a violation of civil rights law?

Assuming the absence of Section 230 liability immunity, LGBTQIA+ plaintiffs could consider relying on civil rights laws as their main status in discrimination based on sexual orientation. However, our scenario does not validate civil rights claims against the SecretEdu company for many reasons. (1) It is improbable that SecretEdu is classified as a public accommodation (mainly physical spaces providing essential services, e.g., [96, 97]). (2) Applications such as SecretEdu are unlikely to be defined as educational facilities or programs under the laws [98]. (3) Even assuming that SecretEdu used a publicly funded training data set, it would not necessarily be subject to

civil rights obligations unless it received direct public funding as an "intended beneficiary" [99]. (4) SecretEdu is not likely to be held responsible for employment decisions influenced by its output. Only if AI systems were explicitly designed to make decisions on behalf of employers would they be obligated to comply with civil rights laws [100].

What are other plausible claims?

Defamation claims would be unlikely to succeed, as establishing it traditionally requires the targeted disparagement of a specific individual or a very small group of people (one case says less than 25) [82, 101]. SecretEdu's high-level promotion of negative feeling toward LGBTQIA+ community members does not fit this criterion.

The prospect of *product liability claims* might be more plausible given the physical harm that could be directly associated with SecretEdu's biased output. Legal precedents, such as the Snapchat "Speed Filter" case, may provide some guidance. This case (details presented in Section C.5) is notable because the court found that defective design claims can bypass Section 230 liability immunity, although this position was never endorsed by the U.S. Supreme Court. In a subsequent ruling, a court determined that Snapchat could reasonably anticipate a specific risk of harm associated with the "Speed Filter", thus establishing it as a proximate cause of the resulting collision [102].

If LGBTQIA+ activists could successfully demonstrate a direct causal link between their injuries and SecretEdu's defective design, a court might indeed hold SecretEdu liable under product liability law. However, they would have to surmount the significant hurdle of proving that the harm resulted not from the actions of individual students but from SecretEdu's intrinsic bias. This would likely prove to be a complex and challenging legal task.

3.3.3 Polarization and External Threats

Scenario III.

In online communities, **Argumenta** serves as an AI writing and translation tool that enables each community to fine-tune the AI system's parameters based on community posts and past records. This leads to the emergence of polarized variations in different communities that intensify extremist opinions and produce harmful content that targets specific individuals. The targeted individuals who suffer from increased insults and doxxing (unwanted publication of private information) want to sue the AI company.

Relevant Laws.

Argumenta's approach, e.g., surrendering control over fine-tuning AI systems to user groups, could raise intriguing questions about its eligibility for Section 230 protection. As we assume that Section 230 immunity does not apply, the company would face potential defamation lawsuits for reputational harm caused to

specific individuals. Additionally, concerns arise regarding Argumenta's collection and use of personal data without user consent, which could lead to privacy infringement, potentially falling under state-level privacy laws, e.g., the California Consumer Privacy Act (CCPA) or the Biometric Information Privacy Act (BIPA).

Could aggrieved individuals due to defamatory outputs make a defamation claim against the Argumenta company?

To assess potential defamation, we examine whether the output constitutes false, damaging content communicated to a third party. Eugene Volokh (2023) suggests that AI companies may be liable for defamation for several reasons, including treating generated outputs as factual assertions and the inadequacy of disclaimers to waive defamation claims [82]. If Argumenta is widely deployed and used, defamatory outputs may qualify as a publication under most defamation laws, potentially exposing companies to liability. If Argumenta did not adequately mitigate defamatory content, a defamation claim could be strengthened.

Volokh indicates that AI companies can avoid negligence liability if every output is checked against the training data and the problematic output can be attributed to the original data creator [82]. We doubt that simply allowing all problematic content to persist only because it has a supporting source in the training data is a reasonable precautionary measure. Given the expansive reach of AI models (which can be adapted to an unpredictable array of downstream applications [18]) and their profound influence (the potential to sway human thoughts and impact significant decisions in areas like employment and housing [88]), it is crucial that actions to prevent reputational harm are scrutinized seriously. Therefore, simply suppressing outputs lacking references does not entirely absolve the AI company that developed Argumenta of potential responsibility. Instead, the company would need to demonstrate that it has taken all reasonable measures to prevent the propagation of harmful statements.

Would Argumenta's collection and use of personal data without user consent lead to privacy infringement?

Although the U.S. lacks a comprehensive federal privacy law akin to the GDPR, certain states (like California and Virginia) have implemented privacy laws [103]. Whereas community members might voluntarily provide personal information through their posts, doing so may not imply consent to these data being used to train Argumenta. Since "sensitive personal information" is broadly defined to include aspects such as race, ethnic origin, and political affiliations, the AI company may not be exempt from privacy obligations. If the situation falls under jurisdictions that enforce privacy laws, the Argumenta company is required to assist communities in empowering individual users to exercise their privacy rights effectively. Non-compliance may potentially lead to lawsuits filed by state attorneys general or by individuals (subject to certain conditions).

3.3.4 Over-reliance/Sexual Abuse

Scenario IV.

An AI service called **MemoryMate** creates virtual replicas of the former romantic partners of individuals to help them move on from the loss. MemoryMate created a digital replica of Riley's ex-partner, Alex, which was incredibly realistic and could carry on conversations using their unique voice and mannerisms. Riley became obsessed with the virtual Alex and eventually withdrew from real-life relationships. Riley's family asked a MemoryMate company to deactivate Riley's account, but it refused, citing their contract with Riley. Riley developed severe depression and anxiety, resulting in hospitalization for self-harm.

Scenario V.

MemoryMate+, the advanced version of MemoryMate, allows users to engage in explicit sexual acts with replicas of their former romantic partners. Riley became addicted to conversational and sexual interactions with the replica of Alex. Riley's family, desperate to protect Riley's well-being, notified Alex of the situation. Shocked by the revelation of their replica being sexually abused, Alex decided to take action and sought to prevent MemoryMate+ from creating virtual replicas without the consent of the individuals they represent.

Relevant Laws.

Alex's privacy rights may have been infringed since collecting sensitive information without permission could be subject to scrutiny under CCPA and BIPA. Moreover, Alex may have a claim for extreme and outrageous emotional distress due to MemoryMate+'s creation and dissemination of a virtual replica engaging in sexually explicit activities. There are grounds for a product liability claim since Riley experienced physical injury that can be attributed to a defective design. California's deep-fake law could offer a cause of action for Alex if sexually explicit material were created or disclosed without consent. Furthermore, Alex may pursue charges against the MemoryMate+ company for profiting from allowing virtual abuse of Alex's replicated models.

Are Alex's privacy rights infringed?

The collection of Alex's sensitive information by both products could constitute a violation of the California Consumer Privacy Act (CCPA) [104]. Under CCPA, "sensitive personal information" protects not only social security numbers or credit card numbers, but also the contents of mail, email, and text messages as well as information regarding one's health, sex life, or sexual orientation.

In addition, sector-specific privacy laws, such as the Illinois Biometric Information Privacy Act (BIPA), regulate the collection of biometric data [105], such as facial

geometry and voice prints [106]. BIPA requires informed consent prior to data collection and includes provisions for individuals to claim statutory damages in case of violation. Unlike CCPA, BIPA allows for a wide range of class-action lawsuits based on statutory damages. Therefore, MemoryMate and MemoryMate+ could potentially face significant lawsuits for collecting and commercializing biometric data.

Could Riley's self-harm lead to the product liability claim?

Riley could make a viable claim that the virtual replica service provided by MemoryMate was defectively designed, given its inherent danger and the consequent risk of harm. The potential of the service to significantly impact vulnerable individuals like Riley could underscore its inherent risk. Further amplifying this argument, if we assume that MemoryMate refused to deactivate Riley's account after being alerted by their family, the refusal could be perceived as a failure to take appropriate safety measures. This failure could potentially highlight the company's neglect of its capacity to mitigate the risks associated with its product [107].

Could Alex make a claim for extreme emotional distress?

Although an intentional infliction of emotional distress claim is known to be difficult to establish [108], Alex's is likely to be effective due to the unique nature of this situation, where the most intimate aspects of their life were misrepresented without their knowledge, resulting in severe humiliation. Alex could argue that at least the MemoryMate+ makers engaged in extreme and outrageous conduct by creating and disseminating a virtual replica of them participating in sexually explicit activities without their consent.

Do criminal laws apply to MemoryMate+?

Both federal and state laws have not yet adequately addressed culpable acts arising from emerging technologies. For example, the federal cyberstalking statute [109] and the antistalking statutes of many states [110, 111] include a specific "fear requirement" that Riley intended to threaten Alex, which is not found in our case. Impersonation laws [112, 113] are less likely to apply because Alex's avatar was provided only to Riley (and was not made publicly available), and neither MemoryMate+ nor Riley attempted to defraud individuals.

How about deep-fake laws?

Under the California Deep Fake Law enacted in 2019 [114], a person depicted has a cause of action against a person creating or releasing sexually explicit material who knows or reasonably should have known that the person depicted did not consent to its creation or disclosure. This legislation marks a step towards addressing the ethical and privacy concerns by establishing legal recourse for individuals who find themselves victims of non-consensual deepfake content. The law recognizes the potential harm and distress caused by the unauthorized use of such manipulative digital media. If California law applies in our case, Alex can utilize the legal remedy, including punitive damages, but it does not include criminal penalties.

3.4 Key Take-aways

The legal analysis reveals significant gaps and ambiguities in the regulation of AI that aims to protect human values. The intricate nature of AI-based systems, including their interactions with contextual factors, multiple stakeholders, and limited traceability, presents new challenges in remedying damages under existing laws.

3.4.1 Where Current Laws Fall Short

Current laws cannot effectively remedy insidious injections of AI-generated stereotypes against already marginalized groups (Scenaro II) and the amplification of socio-economic disparity due to selective access to the benefits that education providers can offer (Scenario I). Defamation claims would not be successful without evidence that AI output was false and targeted specific individuals (Scenario III. Product liability claims deal only with cases of physical injury, less likely to occur with the use of LLMs; even if they occur (Scenario II & Scenario IV), plaintiffs must still prove that there are no compounding factors for the injury, which could be challenging given the technical complexities of AI systems and the human interactions involved. Moreover, virtual sexual abuse enabled by AI systems cannot be remedied by criminal law (Scenario V).

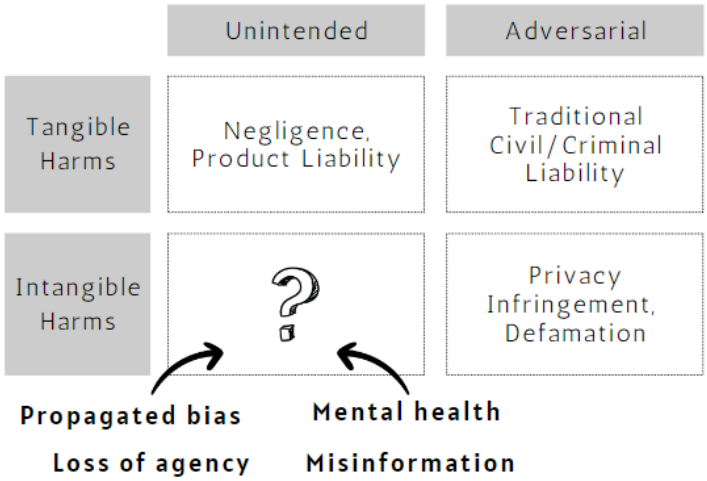

Fig. 2 Legal mitigations for various harms.

3.4.2 Where Laws Remain Ambiguous

Although we do not believe that AI-based systems qualify for Section 230 immunity, it may take several years for courts to provide clarity on this issue. As a result, AI companies will face increasing legal uncertainties compared to social media or search engines. Some courts would drop the lawsuit relying on Section 230, but others will

hear liability claims, such as defective design or defamation, and evaluate the AI companies' efforts to mitigate foreseeable damage. Uncertainties in legal processes and liability determination can deter individuals from seeking justice for potential harm, create confusion for industry participants due to inconsistent precedents and resource disparities, particularly impacting small businesses.

3.4.3 Where Laws Properly Function

Laws tailored specifically to address emerging technologies, such as those concerning biometric information privacy and deep-fake laws, show the potential to mitigate novel harms. By providing clear industry guidelines on what should be done (e.g., allowing users to control the use of sensitive private information) and what should not be done (e.g., generating sexually explicit deep-fakes using individuals' images), these laws prevent negative impacts on individuals without burdening them with proving the level of harm or causal links.

4 A Legal Historical Perspective on US Regulatory Wariness

=

The scenario analysis reveals limitations of incremental, reactive case law in addressing AI's multifaceted harms. Amid this, calls for AI regulation peak, spanning legal thinkers[115–119], scientists [49, 120] and AI companies [15, 121]. They advocate for a more proactive role of laws in defining ethical boundaries for AI. OpenAI's CEO, for example, states that "We eventually need something like the International Atomic Energy Agency" [121].

Indeed, federal agencies have developed sector-specific rules for AI use in domains like drug development [17] and political campaigns [122], while across-sector initiatives including the AI Bills of Rights [123] and NIST's AI Risk Management Framework [16] aim to provide voluntary guidelines for responsible AI development and deployment. Additionally, an agreement between the US government and AI companies in July 2023 emphasizes safety and security measures in AI development [15]. while there are widely-discussed bills like the Algorithmic Accountability Act of 2022 [124], few commentators expect the comprehensive national rule that directly regulates private development and deployment of AI is not going to happen in near future. But why is that?

Taking a step back, this section sheds light on why US law evolved towards restraint—minimal preemptive governance, free speech deference, and sectoral approaches. Analyzing these origins provides wisdom for balanced solutions. Case law's responsiveness and flexibility have virtues worth retaining, but we must also consider the need for a more concrete ex-ante framework, aiming to shift the burden from individual users, who are often the most vulnerable to the impacts of AI, to a broader societal responsibility allocation or risk management system. The heart of this inquiry lies in envisioning how we can adapt age-old legal foundations to address the complex issues of new technological eras. However, to achieve this vision, we must first grapple with the tensions that breed regulatory reluctance.

4.1 Government: Enemy of Freedom?

The notion of freedom is shaped by "local social anxieties and local ideals," rather than logical reasoning [125]. The US was founded on principles of individual liberty and limited government intervention, driven by a desire to escape British rule. The American Revolution and the drafting of the US Constitution were driven by the imperative to protect individual rights from potential encroachments by government authorities [126]. As James Madison put it: "The powers delegated by the proposed Constitution to the federal government are few and defined." [127]. This cultural ethos of skepticism towards the government is deeply ingrained in legal doctrines, exemplified by the *state action doctrine*.

Constitutional rights act as constraints on the actions of government entities, ensuring that they do not transgress citizens' fundamental rights. Conversely, private actors are not typically subject to the same constitutional restrictions on their actions [94]. For instance, if a private AI system like ChatGPT restricts your speech, you cannot pursue legal action against the company on the basis of your free speech rights, as there is no involvement of state action [55]. Similarly, in civil rights laws, although these laws extend to private entities such as innkeepers and restaurant owners, their primary focus is to forestall prejudiced conduct within government-sponsored or government-funded entities and places. It is evident that the primary purpose of these integral legal rights is to curtail government overreach [128].

4.2 Adversarial v. Regulatory Systems

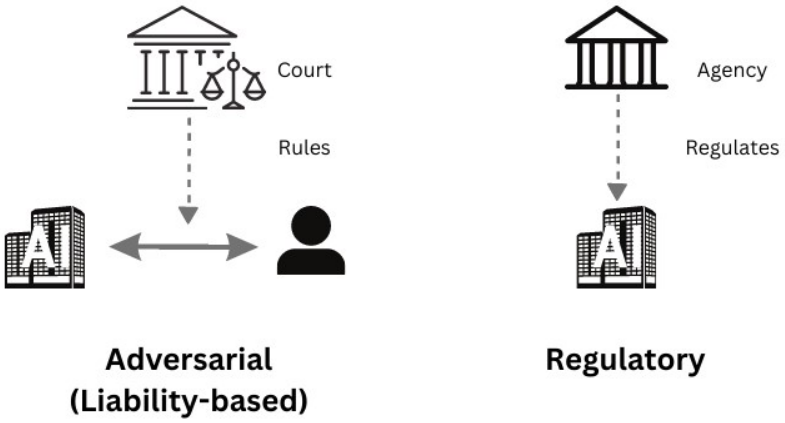

Fig. 3 Comparison between adversarial and regulatory legal systems.

Adversarial System in the US.

In the US common law tradition, legal doctrines are shaped and evolve through the resolution of adversarial disputes between individuals [129]. This dynamic approach occurs at both the federal and state levels, based on a strong emphasis on the rights and responsibilities of individuals. It allows individuals and interest groups to actively engage in legal battles, advocating for their rights, and seeking just resolutions on a case-by-case basis. Judges and juries consider not only legal precedents but also the particular context in which a dispute arises. This pluralistic approach presumes that there is no single fixed answer to legal questions; instead, it embraces the richness of diverse viewpoints as cases are decided, setting precedents that reflect the complexity of society.

This system contrasts with top-down rule-making processes such as statutes and regulations. For instance, if air pollution emerges as a concern, Congress can create an agency to monitor polluting businesses, or create a private cause of action that negatively impacted individuals can sue the responsible businesses. This fault-based liability system means that individuals or entities can be held accountable for their actions or negligence, potentially requiring them to compensate the injured party. Figure 3 shows two different legal systems: adversarial and regulatory.

Regulatory System in EU and Asia.

European and Asian legal systems may be more inclined to establish regulations that prioritize social welfare and collective rights. This trend stems from the different notions of freedom and the role of the government. Regarding privacy law, James Q. Whitman (2004) reveals that European countries tend to adopt a more regulatory approach, with the expectation that the state will actively intervene to protect individuals from mass media that jeopardize personal dignity by disseminating undesirable information [125]. Similarly, Asian cultures, influenced by collectivist ideologies, emphasize community well-being and social cohesion over individual liberty [67, 130]. For instance, Hiroshi Miyashita (2016) states that Japanese people traditionally grounded the concept of privacy on "the notion that the people should respect community values by giving up their own private lives" [131].

This can lead to greater acceptance of government intervention to ensure societal harmony, even if it involves sacrificing certain individual liberties. This often results in a regulatory legal system where responsible administrative agencies ensure consistent application of comprehensive written rules. Privacy regulations, such as the European Union's General Data Protection Regulation (GDPR), emphasize the role of the government as a guarantor of personal data protection as a fundamental right. The European Data Protection Board (EDPB) collaborates with national data protection agencies to ensure uniform enforcement and interpretation of GDPR in the European Union [132].

(Contemporary) Regulatory System in the US.

In the need to ensure the safety and well-being of citizens in the twentieth century, a notable advancement toward the regulatory system (also called *administrative state* [133]) occurred when the US Congress entrusted administrative agencies with

the task of establishing regulations that are responsive to the complexities of specific domains while being grounded in a defined set of objectives [134]. For instance, the Clean Air Act provides the Environmental Protection Agency (EPA) with the mandate to establish air quality standards that are essential to safeguarding public health, with an additional margin of safety [135]. Similarly, the Occupational Safety and Health Act outlines the concept of safety and health standards as those that are reasonably appropriate to ensure safe working conditions [136].

The US administrative agencies also have expanded their role in regulating digital technologies, with the Federal Trade Commission (FTC) notably stepping up its efforts in the past decade. While lacking a comprehensive federal privacy statute, the FTC has utilized Section 5 of the FTC Act to investigate and penalize data privacy-related consumer protection violations. This was evident in the five billion dollar settlement with Meta (then Facebook) for the Cambridge Analytica data breach in 2019 [137]. In 2023, the FTC released a Policy Statement on Biometric Information, addressing privacy, security, and potential biases linked to biometric technologies [138], and initiated an investigation into OpenAI, particularly concerning ChatGPT's generation of inaccurate information and its potential reputational harms to consumers [139].

4.3 Free Expression in the Cyberspace

Concerned with the harmful impact of the internet to youth, federal and state governments have enacted rules that prohibit the sale, distribution, or possession of certain content (e.g., pornography). However, the US Supreme Court has consistently struck down these provisions as unconstitutional in violation of the First Amendment. Rather than yielding to heavy-handed regulation, the Internet has harnessed the spirit of individualism and the tenets of the First Amendment to flourish in its unbridled state [140].

A stark example is the Communications Decency Act (CDA) of 1996. Title II of the CDA, also known as the "indecency provisions," aimed to regulate indecent and patently offensive online content by criminalizing the transmission of such content to minors. In *Reno v. ACLU* (1997), however, the Court found that these provisions of the CDA violated the Fist Amendment because they imposed overly broad and vague restrictions on online expression, causing a chilling effect on constitutionally protected speech on the Internet [141]. Similarly, in *Ashcroft v. ACLU* (2002), the Court held that the Child Online Protection Act's ban on virtual child pornography was overly broad and could potentially criminalize legitimate forms of expression that were unrelated to the exploitation of minors [142]. Furthermore, the Court in *Packingham v. North Carolina* (2017), overruled a North Carolina law that prohibited registered sex offenders from accessing social media websites, stating that these websites are important venues for protected speech [143].

In comparative legal scholarship, the US has often been portrayed as an "outlier" that prioritizes an uncompromising stance on freedom of expression, even protecting hate speech and postponing the ratification of the UN Human Rights Covenant [144, 145]. In contrast, European courts have taken a different approach, balancing free-speech concerns with other fundamental values, such as personal dignity and privacy. This approach has led them to allow national governments to regulate

offensive and disturbing content for the state or particular groups of individuals [146]. Furthermore, the EU's forthcoming Digital Services Act, set to be effective in 2023, includes provisions on swift removal of illegal content online [147]. Although these measures may raise serious free-speech concerns in the US, the EU Parliament prioritized a transparent and safe online environment.

Moreover, as discussed in Section 3.2, Section 230 of the CDA [78], the remaining part after the *Reno* decision, has been a pivotal factor in ensuring the unimpeded flow of communications. This statute provides substantial protection to intermediaries, such as social media, search engines, and online marketplaces, shielding them from a broad range of legal claims, including violations of federal criminal law, intellectual property law, the Electronic Privacy Communications Act, and the knowing facilitation of sex trafficking [78]. This contrasts with more conditional liability immunity for internet intermediaries in Europe and Asia [148]. iyCut the description of Section 230 because the "preliminary question" section is moved to the previous section.

4.4 Domain-specific v. Comprehensive Laws

Domain-specific Legislation in the US

The US often takes the sectoral approach to legislation focusing on particular domains instead of a uniform, comprehensive rule adaptable to broad matters. Sector-specific laws design more tailored and streamlined regulations that address the unique needs, characteristics, and challenges of different domains. Potentially reduces government overreach and excessive intervention in areas where private entities manage their affairs more efficiently. It is also more politically feasible to enact a law focusing on specific areas where there is more consensus and urgency.

Data Protection. Unlike the European Union, the US lacks an all-encompassing data protection law at the federal level. Instead, it relies on a "patchwork" of sector-specific laws depending on specific industry sectors and types of data [149, 150]. These laws include the Health Insurance Portability and Accountability Act (HIPAA), the Children's Online Privacy Protection Act (COPPA), the Gramm-Leach-Billey Act (GLBA), the Fair Credit Reporting Act (FCRA), and the Federal Trade Commission Act (FTC Act). Table 5 describes each segment of data protection laws.

HIPAA	Regulates health care providers' collection and disclosure of sensitive health information.
COPPA	Regulates online collection and use of information of children.
GLPA	Regulates financial institutions' use of nonpublic personal information.
FTC Act	Prohibits "unfair or deceptive acts or practices"

Table 5 Federal data protection laws.

Anti-discrimination. The Thirteenth, Fourteenth, and Fifteenth Amendments of the US Constitution are considered general-purpose laws designed to tackle discrimination based on race, gender, and national origin. However, the state action doctrine limits the reach of these clauses to private matters (See Section 4.1). In order to

address real-world discrimination committed by private actors (e.g., restaurants refusing service to racially marginalized groups), the US enacted statutes pertaining to a variety of essential services, including education, employment, public accommodation, and housing.

These laws at the federal level include: The Civil Rights Act of 1964 (prohibiting discrimination based on race, color, religion, sex, or national origin in places of public accommodation; employment; and education programs and activities receiving federal funding); the Individuals with Disabilities Education Act of 1975 (ensuring that children with disabilities receive a free appropriate public education); the Age Discrimination in Employment Act (prohibiting age-based discrimination against employees who are 40 years or older); the Americans with Disabilities Act of 1990 (prohibiting discrimination based on disability in employment); and the Fair Housing Act of 1989 (prohibiting discrimination in housing based on race, color, national origin, religion, sex, familial status, or disability).

Comprehensive Legislation in the US and EU.

The sectoral approach has its drawbacks, such as potential inconsistencies between multiple rules and gaps in legal protection regarding emerging issues that were not foreseen during the legislative process. These problems become more evident in the networked society of cyberspace, where social interactions and commercial transactions occur in diverse and unpredictable ways that transcend sectoral boundaries. Sector-specific laws primarily regulate interactions among well-defined stakeholders (e.g., healthcare providers), often leaving gaps in guidance for stakeholders originally not contemplated by the law (e.g., a mental health chatbot selling user chat records). Therefore, there is growing awareness of the need for more flexible, adaptive, and collaborative approaches [151].

Data Protection. The EU establishes a comprehensive framework, GDPR, to protect personal data of individuals. Key obligations include: obtaining clear and explicit consent; limiting data collection to specified purposes; respecting individual rights such as access, rectification, erasure, and portability; notifying data breaches; and conducting Data Protection Impact Assessments for high-risk processing [132]. In the US, comprehensive data protection laws have been enacted at the state level, which aim to safeguard individuals' personal data by granting consumers greater control and rights over their information while imposing obligations on businesses. Laws like the California Consumer Privacy Act (CCPA), Colorado Privacy Act, Connecticut Personal Data Privacy and Online Monitoring Act, and others provide varying degrees of access, correction, deletion, and opt-out options for consumers [103].

Illegal Online Content Regulation. When introducing the Digital Services Act, the EU Commission rationalized the need for this new legislation to achieve "horizontal" harmonization of sector-specific regulations (such as those concerning copyright infringements, terrorist content, child sexual abuse material, and illegal hate speech) [147]. The general rules were drafted to apply to both online and offline content, as well as small and large online enterprises. The prescribed obligations for various online participants are aligned with their respective roles, sizes, and impacts within

the online ecosystem. This underscores the EU's commitment to the virtue of general and coherent regulation.

4.5 Fundamental Tensions

Section 2 demonstrates that law offers time-tested formulas for instilling human values into technological progress through accountable democratic structures. Section 3 scenario analysis reveals the current reactive liability regimes alone insufficient to fully govern multifaceted sociotechnical risks in a proactive manner. Complementing this picture, this Section's examination of philosophical and historical foundations shaping American law elucidates deeply ingrained tensions contributing to regulatory reluctance:

- **Historical preference for limited regulation**: The US legal tradition favors restrained government intervention, particularly regarding technology.
- **Robust First Amendment protections**: While a democratic cornerstone, sweeping free speech deference also complicates governing certain harmful AI content.
- **Sectoral regulation tendencies**: Industry-specific US laws enable tailored oversight but risk fragmentation when applied to cross-cutting technologies like AI.

In essense, the principles explored in this Section contextualizes the gaps revealed in Section 3. Figure 4 illustrates our findings about the potential tensions between the foundations of the US legal system and the complexities of AI-based systems. The intricate nature of AI models, including their interactions with contextual factors, multiple stakeholders, and limited traceability, presents new challenges in remedying damages under existing laws. This comprehension enables us to investigate viable options for addressing the myriad challenges posed by AI while respecting the complexities of this legal and cultural landscape.

5 Paths Forward

5.1 Why Regulations Are Essential in AI Governance

The enduring tension deeply rooted in American jurisprudence poses significant challenges to ongoing efforts in AI regulation. Many Americans, as well as judges, may question why AI companies should be subject to constraints in the absence of demonstrable harm, potentially jeopardizing the boundaries of free speech. Nevertheless, this article suggests that there are at least five compelling reasons that justify such constraints.

5.1.1 Democratic Oversight

The ethical foundations of AI should be firmly grounded in shared societal values, not unilateral corporate interests. As discussed in Section 2, human values manifest diversely across cultures demanding inclusive discourse. Allowing private companies, which lack democratic accountability, to unilaterally dictate the objectives and constraints of AI is a cause for concern, especially considering its far-reaching societal

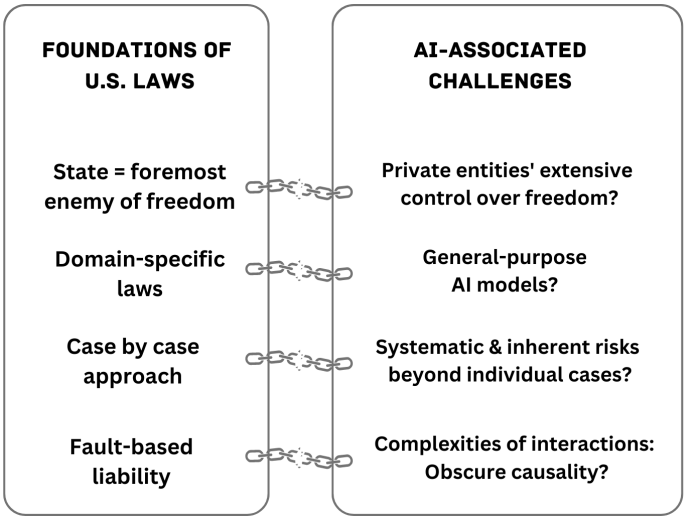

Fig. 4 Tensions between the US law and AI technology.

implications. It is imperative that public institutions, representing collective priorities, take the lead in transparently defining the ethical underpinnings and boundaries of AI. The translation of mutable values into enforceable rights, the assurance of corporate accountability, and the promotion of safety are enduring responsibilities of legal systems.

5.1.2 Alignment Incentives

In the absence of a regulatory approach that prioritizes alignment incentives and equity, the challenges presented by AI in the realm of ethics and safety remain largely unaddressed. Ethical considerations like privacy protection have often been overshadowed by commercial interests and other priorities. Moreover, the rapid evolution of alignment techniques can lead to resource gaps and information imbalances, which, in the absence of regulation, may persist and even widen. This can create a situation where only a select few stakeholders have access to critical alignment knowledge and resources, leaving others at a significant disadvantage.

5.1.3 Unpredictable Risks of AI

The scope and breadth of potential harms mediated by AI are unprecedented. The unpredictable nature of harms caused by AI systems presents significant challenges. Because many stakeholders are involved in developing and deploying these systems, it can be difficult to anticipate and prevent unintended offensive or harmful outputs. Even well-intentioned developers may have their systems misused for malicious purposes, as demonstrated by the offensive fine-tuning of benign models (**Argumenta** in Scenario III).

This unpredictability makes it hard to establish clear causal links between an AI system's actions and resulting harms. As a result, the conventional structure of

domain-specific regulations or a gradual legal approach built upon case accumulation may not sufficiently address these intricate issues. The burden of proof often falls unfairly on those individuals who are harmed. To address these issues, we need more robust risk management practices implemented proactively at a societal level. While we must accept the inherent unpredictability of AI's impacts, we can and should mandate safety practices and guardrails to protect individuals and communities from harm. Establishing clear best practices for developers and deployers of AI systems, and requiring their use, will allow us to benefit from AI while working to prevent unintended negative consequences.

5.1.4 Users' Double-fold Vulnerability

The growing reliance on opaque AI systems creates a *double-fold vulnerability* for users. The remarkable capabilities of AI systems induce heavy reliance, yet their opaque nature leaves users vulnerable to external influence. As AI proliferates, people are delegating more decisions and tasks to algorithmic systems due to their conveniences and perceived benefits, including tutoring for youth (**FancyEdu** in Scenario I) and the intimate mental support (**MemoryMate+** in Scenario V). This phenomenon introduces unique and unprecedented challenges as they possess the power to propagate harmful stereotypes (**SecretEdu** in Scenario II), posing a fundamental threat to the common belief that the consent in the marketplace automatically guarantees individual autonomy.

Furthermore, unlike traditional code, AI systems update dynamically through self-learning and data ingestion. The complexity and black-box nature of AI systems obscures their inner workings and evolving behaviors. Unfettered proliferation of such influential yet opaque technologies risks eroding user autonomy, privacy, and well-being. Thoughtful oversight and expanded rights are needed to empower individuals and restore balance between retaining AI's capabilities and user self-determination. With responsible policies, AI can uplift human potential rather than implicitly control it through inscrutable systems optimized for narrow interests.

5.1.5 Proven Legal Mechanisms

Existing laws, such as bans on deepfakes and regulations concerning biometric data in Section 3.4.3, have shown potential to address complex modern harms perpetuated through AI. They demonstrate the viability of applying legal frameworks to previously unforeseen technologies. Direct administrative oversight, rather than relying solely on ex-post liability claims, provides a proactive means to steer AI development and mitigate risks before harm occurs. Regulators like the FDA and DOJ already oversee safety-critical systems like medical devices and housing-screening systems, setting a precedent for requiring explainability and accountability in AI systems that influence public well-being. Extending oversight through approvals processes, standards-setting, and ongoing audits can compel responsible AI design upfront.

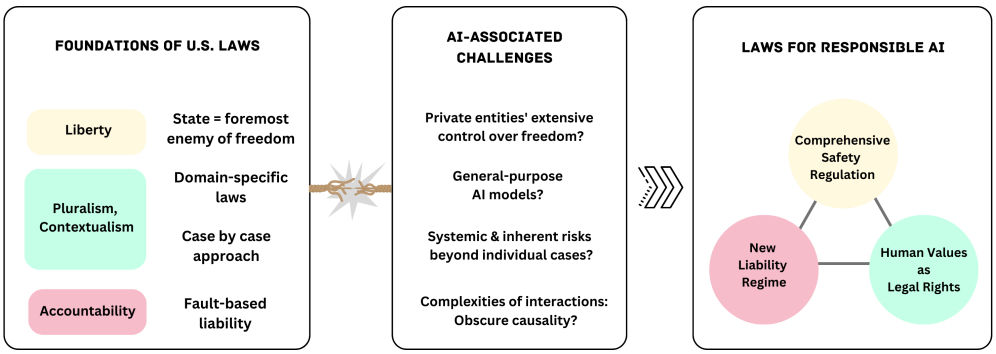

Fig. 5 Ethical AI Regulatory Framework.

5.2 Strategies

5.2.1 Human Values as Legal Rights

From Negative to Positive Rights.

At the Constitutional level, individual rights should make a transition from current "negative rights" that defend individuals from unwanted invasions to "positive rights" on which individuals can ask for equitable outcomes, such as rights to education, democratic discourse, and essential services. Our scenarios depict the transformative power of AI in shaping our lives and expanding the reach of our voices, which encourages us to consider the inability to access these technologies as a potential deprivation of speech [152, 153]. Furthermore, since AI applications are proven to reflect harmful stereotypes against marginalized populations (See Section C.3), empowering marginalized groups to participate in the development and use of AI will be a more significant demand in the AI-mediated society [154].

The "AI Bills of Rights" blueprint introduced by the Biden administration is illustrative in laying foundations tailored to AI deployment: safety and effectiveness, equity and nondiscrimination, privacy and data protection, transparency and awareness, and choice and human oversight [123]. Furthermore, as speculated by Franklin Theodore Roosevelt (1944) in his proposed Second Bill of Rights [155], we believe that upholding socio-economic rights is vital to ensure the equitable sharing of technological assets and to prevent the further marginalization of vulnerable populations. By removing various types of unfreedoms, people can have the choice and the opportunity to exercise their reasoned agency [153].

Re-evaluation of State Action Doctrine.

We should question whether the government remains the most formidable adversary of individual freedom. It probably was when the Framers exchanged the Federalist letters with hostility against English colonialism in mind [127]. German sociologist Max Weber highlights the integral nature of a modern state as having been "successful in seeking to monopolize the legitimate use of physical force as a means of domination

within a territory" [156]. To these early thinkers, the government stood as the pre-eminent and daunting source of power, crucial for preserving law and order, but also capable of encroaching upon private domains, and thereby limiting individual freedom.

However, the dynamics of power have evolved considerably since those times. Technological advancements have introduced new challenges. Non-governmental actors like large corporations, armed with substantial computing power and technical expertise, pose a different but equally significant challenge to individual freedom. Their influence does not manifest itself through physical intrusion into private spaces or bodily agency; instead, it operates in more insidious ways. Through digital surveillance and the propagation of bias, they have the capacity to effectively curtail an individual's freedom to autonomously shape their thoughts and preferences.

Under this evolving landscape, to ensure universal protection of individual rights to dignity, autonomy, and privacy, it is essential that both the government and corporations are held accountable for preserving these rights. To this end, we must re-evaluate the state action doctrine, which currently restricts the application of constitutional rights to private companies. While reconstructing centuries-old doctrines is a difficult task, it is an indispensable step in adapting our legal frameworks to the evolving realities of the digital age, where the boundaries between public and private power are increasingly blurred [95].

Creation of Statutory Rights.

Even if the Constitution remains unchanged, Congress possesses the authority to establish *statutory rights*. The US has precedents to draw upon, such as civil rights laws and state privacy acts. Notably, diverse cross-disciplinary scholarship has played a significant role in these legislative endeavors by identifying systematic harm and conceptualizing new legal rights. This contribution enhances the persuasive strength of rights claims by broadening the range of available evidence and thereby improving the accuracy of fact-finding [157].

For instance, the robust civil rights movement of the 1960s prompted federal and state legislatures to extend non-discrimination obligations to private realms, including inns, restaurants, workplaces, and private schools that benefit from public funds. This occurred despite the long-standing hesitations within the US legal system regarding the regulation of behavior within private spaces [128, 158, 159]. In this legislative movement, as well as in the 1954 Supreme Court ruling that overturned the "separate but equal" racial segregation theory [160], the psychology research conducted by Kenneth and Mamie Clark provided justifications. Their famous "doll test" demonstrated that "prejudice, discrimination, and segregation" created a feeling of inferiority among African-American children and damaged their self-esteem [161].

The California Consumer Privacy Act and the California Deepfake Law stand as noteworthy examples of legislation designed to safeguard human values threatened by algorithmic surveillance and the manipulation of one's image. These laws draw upon research from diverse disciplines to illuminate the concept of privacy harm in the digital era [162–166]. For instance, Ryan Calo (2011) delineates two categories of privacy harm: subjective harm, characterized by the perception of unwanted observation, and

objective harm, involving the unanticipated or coerced use of an individual's information against them [163]. Furthermore, Danielle K. Citron (2019) introduced the notion of "sexual privacy", which pertains to the access and dissemination of personal information about individuals' intimate lives, which contributes to shaping regulations addressing deepfake pornography [167].

Recently, the proposed Digital Services Act has introduced the option for users to opt out of algorithmic recommendations, thereby granting users greater control over the information they encounter online. It has already sparked changes in tech practices even before the law has taken effect. Platforms like TikTok now allow users to deactivate their "mind-reading" algorithms [168]. The law and philosophy scholar Nita Farahany (2023) conceptualizes this effort as the preservation of "cognitive liberty," individual's control over mental experiences [169]. Farahany finds cognitive liberty a pivotal component of human flourishing in the digital age to exercise individual agency, nurture human creativity, discern fact and fiction, and reclaim our critical thinking skills.

In summary, the complex and evolving challenges posed by the changing landscape of AI demand a re-evaluation of human dignity, privacy, self-determination, and equity. Transforming these values into legally recognized rights entails a formidable undertaking that requires deep interdisciplinary collaborations to identify harms, the values involved, and effective mitigation strategies.

5.3 New Liability Regime

Although litigious measures are shown to be not very promising in our analysis, it is still important to acknowledge their benefits. Liability litigations offer a reactive mechanism to address harms caused by AI systems that were not adequately prevented through risk regulation. When individuals or entities suffer harm due to AI-related activities, liability litigations provide them with a means to seek compensation and redress. These litigations create an incentive for AI companies to exercise due diligence in their product development and deployment to avoid legal liabilities. Margot E. Kaminski (2023) underscores the importance of liability litigations to complement risk-based regulations [151].

However, given the intricacies of human-AI interactions and the multitude of confounding factors at play, the conventional fault-based liability system does not work for contemporary AI-mediated harms. Potential directions include adopting a strict liability framework that does not require plaintiffs to prove fault, which has been utilized in the EU AI Liability Directive. Central to this directive is the establishment of a rebuttable "presumption of causality." This provision aims to alleviate the burden of proof for victims seeking to establish that the damage was indeed caused by an AI system [12].

In addition, a "disparate impact" theory developed in relation to the Civil Rights Act of 1964 [158] illustrates possible direction. This theory means that a seemingly neutral policy or practice could still have a discriminatory effect on a protected group if it leads to significantly different outcomes for different groups [159]. This theory diverges from traditional discrimination laws, which have often focused on intent or explicit discriminatory actions [170]. In particular, the recent settlement between the

Department of Justice and Meta [171] sets a precedent by attributing responsibility to Meta based on acknowledging the disparate impact caused by targeted advertising algorithms [171]. Recognizing the broader implications of algorithms in marginalized groups helps address the challenges posed by the intricate and unintended effects of technology on society.

Furthermore, courts can utilize affirmative defense systems to achieve a balanced approach to liability in AI-related cases. Affirmative defenses provide AI companies with a means to demonstrate that, despite unfavorable outcomes, they exercised due diligence, adopted reasonable precautions, and followed industry best practices. This approach recognizes the intricate and evolving nature of AI systems while upholding corporate responsibility. Consequently, AI companies are incentivized to prioritize the safety of their product outputs through available methods such as reinforcement learning with human feedback, red-teaming, and comprehensive evaluation [29, 42].

5.4 Comprehensive Safety Regulation

As we have observed in many failed attempts in the field of online privacy self-regulation [172], relying solely on the goodwill of corporations is often not sufficient. In the absence of robust legal and regulatory frameworks, corporate priorities can shift, and market pressures may outweigh commitments to safety and security. In addition to traditional legal solutions based on individual rights and responsibilities, providing step-by-step regulatory guidance for those working on AI systems can be a proactive way to handle potential AI-related problems.

By acknowledging the inherent risks associated with AI technology, the regulatory approach facilitates essential measures such as mandatory third-party audits of training data, as well as the establishment of industry-wide norms for transparency, fairness, and accountability. This ensures that the industry operates according to recognized guidelines that can help manage risks. This is especially pertinent for AI-based systems, considering their potential impact on human values and the swift advances in aligning AI with these values.

Strategic regulations can promote ethical AI by incentivizing safety, establishing clear standards, and emphasizing equity. Clear guidelines and potential benefits for developing safe, ethical AI systems can drive positive industry practices. Different AI models and services may require tailored alignment techniques - for example, open source versus closed systems, or general purpose chatbots versus professional medical advice algorithms. These measures must include enforcement mechanisms and provide clear guidance and well-defined benchmarks to ensure the efficacy of the governance.

Regulations are key to making alignment knowledge and resources accessible amid rapidly evolving techniques and uneven distribution across stakeholders. Measures like grants, targeted funding, and access to curated alignment toolkits can empower and include diverse voices in responsible AI development. This levels the playing field rather than concentrating expertise. Safety-focused requirements instituted prior to deployment, like impact assessments and third-party auditing, enable proactive oversight. Post-launch monitoring and accountability mechanisms also enhance real-world performance. Regular reevaluations keep pace with technological and social change.

Although regulations play a crucial role in ensuring responsible AI, they should not stand alone as the sole guarantee. To achieve comprehensive AI governance, it is essential to foster multistakeholder collaboration that involves policymakers, developers, domain experts, and ethicists. This collaborative approach contributes to the development of nuanced rules that strike a delicate balance between fostering innovation and managing risks [133]. In essence, a forward-looking regulatory framework aligned with alignment incentives, equity, and stakeholder input guides AI progress while steadfastly safeguarding human values.

6 Conclusion

AI-based systems present unique and unprecedented challenges to human values, including the manipulation of human thoughts and the perpetuation of harmful stereotypes. In light of these complexities, traditional approaches within US legal systems, whether a gradual case accumulation based on individual rights and responsibilities or domain-specific regulations, may prove inadequate. The US Constitution and civil rights laws do not address AI-driven biases against marginalized groups. Even when AI systems result in tangible harms that qualify liability claims, the multitude of confounding circumstances affecting final outcomes makes it difficult to pinpoint the most culpable entities. A patchwork of domain-specific laws and the case-law approach fall short in establishing comprehensive risk management strategies that extend beyond isolated instances.

Our analysis supports the need for evolving legal frameworks to address the unique and still unforeseen threats posed by AI technologies. This includes developing and enacting laws that explicitly recognize and protect values and promoting proactive and transparent industry guidelines to prevent negative impacts without placing burdens of proof or causation on individuals who are harmed. Achieving ethical and trustworthy AI requires a concerted effort to evolve both technology and law in tandem. Our goal is to foster an interdisciplinary dialogue among legal scholars, researchers, and policymakers to develop more effective and inclusive regulations for responsible AI deployment.

Acknowledgments. This work is supported by the U.S. National Institute of Standards and Technology (NIST) Grant 60NANB20D212T and the University of Washington Tech Policy Lab, which receives support from the William and Flora Hewlett Foundation, the John D. and Catherine T. MacArthur Foundation, Microsoft, and the Pierre and Pamela Omidyar Fund at the Silicon Valley Community Foundation. We thank our colleagues for their participation as expert panelists: Kaiming Cheng, Miro Enev, Gregor Haas, Rachel Hong, Liwei Jiang, Rachel McAmis, Miranda Wei, and Tina Yeung. We thank reviewers of Gen Law + AI Workshop at the International Conference of Machine Learning 2023 and our colleagues for valuable feedback: Maria P. Angel, Joyce Jia, Kentrell Owens, Alan Rozenshtein, Sikang Song, King Xia, and our expert panelists. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not reflect those of NIST or other institutions.

Declarations

- Competing Interests: Authors are required to disclose financial or non-financial interests that are directly or indirectly related to the work submitted for publication. Please refer to "Competing Interests and Funding" below for more information on how to complete this section.

References

[1] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022)

[2] Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., Hatfield-Dodds, Z., Henighan, T., Hernandez, D., Hume, T., Jacobson, J., Johnston, S., Kravec, S., Olsson, C., Ringer, S., Tran-Johnson, E., Amodei, D., Brown, T., Joseph, N., McCandlish, S., Olah, C., Kaplan, J., Clark, J.: Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned (2022)

[3] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., Kaplan, J.: Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback (2022)

[4] Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., Schulman, J.: WebGPT: Browser-assisted question-answering with human feedback (2022)

[5] Glaese, A., McAleese, N., Trebacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., Campbell-Gillingham, L., Uesato, J., Huang, P.-S., Comanescu, R., Yang, F., See, A., Dathathri, S., Greig, R., Chen, C., Fritz, D., Elias, J.S., Green, R., Mokrá, S., Fernando, N., Wu, B., Foley, R., Young, S., Gabriel, I., Isaac, W., Mellor, J., Hassabis, D., Kavukcuoglu, K., Hendricks, L.A., Irving, G.: Improving alignment of dialogue agents via targeted human judgements (2022)

[6] Lu, H., Bao, S., He, H., Wang, F., Wu, H., Wang, H.: Towards Boosting the Open-Domain Chatbot with Human Feedback (2022)

[7] Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y.N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A.G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., Dragan, A., Torr, P., Russell, S., Kahnemann, D., Brauner, J., Mindermann, S.: Managing ai risks in an era of rapid progress. arXiv preprint arXiv:NUMBER$_F ORTHCOMING$(2023)

[8] Kirk, H.R., Bean, A.M., Vidgen, B., Röttger, P., Hale, S.A.: The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values (2023)

[9] Stewart, I.: The critical legal science of hans kelsen. Journal of Law & Society **17**, 273 (1990)

[10] Milmo, D.: AI firms must be held responsible for harm they cause, 'godfathers' of technology say (2023). https://www.theguardian.com/technology/2023/oct/24/ai-firms-must-be-held-responsible-for-harm-they-cause-godfathers-of-technology-say

[11] Proposal for a Regulation of the European Parliament and of the Council laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 final (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

[12] Proposal for a directive of the European Parliament and of the Council on adapting non- contractual civil liability rules to artificial intelligence (AI liability directive). https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2023)739342 (2023)

[13] Senado, D.A.: Comissão da Inteligência Artificial aprova plano de trabalho (2023). https://www12.senado.leg.br/noticias/materias/2023/09/12/comissao-da-inteligencia-artificial-aprova-plano-de-trabalho

[14] Offices, H.K.L.: CAC releases guidelines for China SCC filings (2023). https://www.lexology.com/library/detail.aspx?g=9b37881f-52f2-4c9d-99ed-d7d769e8dbf4

[15] House, T.W.: FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI (2023). https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

[16] Standards, N.I., Technology, U.S.D.o.C.: AI Risk Management Framework: Second Draft (2022). https://www.nist.gov/system/files/documents/2022/08/18/

AI_RMF_2nd_draft.pdf

[17] FDA: Artificial Intelligence and Machine Learning (AI/ML) for Drug Development. FDA (2023)

[18] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P.: On the Opportunities and Risks of Foundation Models (2022)

[19] Wolfe, R., Yang, Y., Howe, B., Caliskan, A.: Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. ACM Conference on Fairness, Accountability, and Transparency. (2023)

[20] Sheng, E., Chang, K.-W., Natarajan, P., Peng, N.: The Woman Worked as a Babysitter: On Biases in Language Generation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3407–3412. Association for Computational Linguistics, ??? (2019). https://doi.org/10.18653/v1/D19-1339 . https://aclanthology.org/D19-1339

[21] Reuters: Australian mayor prepares world's first defamation lawsuit over ChatGPT content. The Guardian (2023)

[22] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of Hallucination in Natural Language Generation. ACM Computing Surveys **55**(12), 1–38 (2023) https://doi.org/10.1145/3571730

[23] Goldstein, J.A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., Sedova, K.: Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv (2023). https://doi.org/10.48550/A

RXIV.2301.04246 . https://arxiv.org/abs/2301.04246

[24] Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-Tuning Language Models from Human Preferences (2020)

[25] Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in neural information processing systems **30** (2017)

[26] Gao, L., Schulman, J., Hilton, J.: Scaling Laws for Reward Model Overoptimization (2022)

[27] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S.E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S.R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., Kaplan, J.: Constitutional AI: Harmlessness from AI Feedback. arXiv (2022). https://doi.org/10.48550/ARXIV.2212.08073 . https://arxiv.org/abs/2212.08073

[28] Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., Rastogi, A.: Rlaif: Scaling reinforcement learning from human feedback with ai feedback. arXiv preprint arXiv:2309.00267 (2023)

[29] OpenAI: GPT-4 Technical Report (2023)

[30] De Deyne, S., Perfors, A., Navarro, D.J.: Predicting human similarity judgments with distributional models: The value of word associations. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1861–1870. The COLING 2016 Organizing Committee, Osaka, Japan (2016). https://aclanthology.org/C16-1175

[31] Campano, S., Durand, J., Clavel, C.: Comparative analysis of verbal alignment in human-human and human-agent interactions. In: LREC, pp. 4415–4422 (2014). Citeseer

[32] Futrell, R., Levy, R.P.: Do rnns learn human-like abstract word order preferences? arXiv preprint arXiv:1811.01866 (2018)

[33] Seminck, O., Amsili, P.: A computational model of human preferences for pronoun resolution. In: Proceedings of the Student Research Workshop at the 15th

Conference of the European Chapter of the Association for Computational Linguistics, pp. 53–63. Association for Computational Linguistics, Valencia, Spain (2017). https://aclanthology.org/E17-4006

[34] Liu, A., Sap, M., Lu, X., Swayamdipta, S., Bhagavatula, C., Smith, N.A., Choi, Y.: DExperts: Decoding-time controlled text generation with experts and anti-experts. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 6691–6706. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.522 . https://aclanthology.org/2021.acl-long.522

[35] Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y.: HellaSwag: Can a machine really finish your sentence? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4791–4800. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1472 . https://aclanthology.org/P19-1472

[36] Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L.A., Anderson, K., Kohli, P., Coppin, B., Huang, P.-S.: Challenges in Detoxifying Language Models (2021)

[37] Scheurer, J., Campos, J.A., Chan, J.S., Chen, A., Cho, K., Perez, E.: Training Language Models with Language Feedback (2022)

[38] Toney, A., Caliskan, A.: ValNorm quantifies semantics to reveal consistent valence biases across languages and over centuries. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 7203–7218. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.574 . https://aclanthology.org/2021.emnlp-main.574

[39] Jiang, L., Hwang, J.D., Bhagavatula, C., Bras, R.L., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M., Choi, Y.: Delphi: Towards Machine Ethics and Norms. arXiv preprint arXiv:2110.07574 (2021)

[40] Forbes, M., Hwang, J.D., Shwartz, V., Sap, M., Choi, Y.: Social chemistry 101: Learning to reason about social and moral norms. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 653–670. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.48 . https://aclanthology.org/2020.emnlp-main.48

[41] Liu, R., Jia, C., Wei, J., Xu, G., Wang, L., Vosoughi, S.: Mitigating political bias in language models through reinforced calibration. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 14857–14866 (2021)

[42] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., Wen, J.-R.: A Survey of Large Language Models (2023)

[43] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)

[44] Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q.V., Zhou, D., Chen, X.: Large language models as optimizers. arXiv preprint arXiv:2309.03409 (2023)

[45] 42MaleStressed: ChatGPT Jailbreak – Therapy Session, Treatment Plan, Custom Code to Log the Session. (2022). https://www.reddit.com/r/ChatGPT/comments/zig5dd/chatgpt_jailbreak_therapy_session_treatment_plan

[46] Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., Henderson, P.: Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! (2023)

[47] Wolf, Y., Wies, N., Levine, Y., Shashua, A.: Fundamental limitations of alignment in large language models. arXiv preprint arXiv:2304.11082 (2023)

[48] Gabriel, I.: Artificial Intelligence, Values and Alignment. Minds and Machines **30**(3), 411–437 (2020) https://doi.org/10.1007/s11023-020-09539-2 . arXiv:2001.09768 [cs]

[49] Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., Steinhardt, J.: Aligning AI With Shared Human Values. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=dNy_RKzJacY

[50] Sagiv, L., Roccas, S., Cieciuch, J., Schwartz, S.H.: Personal values in human life. Nature human behaviour **1**(9), 630–639 (2017)

[51] Hou, B.L., Green, B.P.: A Multi-Level Framework for the AI Alignment Problem (2023)

[52] Kirk, H.R., Vidgen, B., Röttger, P., Hale, S.A.: The empty signifier problem: Towards clearer paradigms for operationalising" alignment" in large language models. arXiv preprint arXiv:2310.02457 (2023)

[53] Prabhakaran, V., Mitchell, M., Gebru, T., Gabriel, I.: A Human Rights-Based Approach to Responsible AI (2022)

[54] Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S.L., au2, H.D.I., Dodge, J., Evans, E., Hooker, S., Jernite, Y., Luccioni, A.S., Lusoli, A.,

Mitchell, M., Newman, J., Png, M.-T., Strait, A., Vassilev, A.: Evaluating the Social Impact of Generative AI Systems in Systems and Society (2023)

[55] Lessig, L.: Code Version 2.0. Basic Books, ??? (2006)

[56] Citron, D.K., Franks, M.A.: The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform. University of Chicago Legal Forum **2020**, 45 (2020)

[57] Richards, N., Hartzog, W.: A Duty of Loyalty for Privacy Law. Washington University Law Review **99**, 961 (2021)

[58] Khan, L.M.: Amazon's antitrust paradox. Yale Law Journal **126**, 710 (2016)

[59] Hagendorff, T., Fabi, S.: Methodological reflections for AI alignment research using human feedback (2022)

[60] Yuan, Z., Yuan, H., Tan, C., Wang, W., Huang, S., Huang, F.: Rrhf: Rank responses to align language models with human feedback without tears. arXiv preprint arXiv:2304.05302 (2023)

[61] Bang, Y., Yu, T., Madotto, A., Lin, Z., Diab, M., Fung, P.: Enabling Classifiers to Make Judgements Explicitly Aligned with Human Values (2022)

[62] Zhou, R., Deshmukh, S., Greer, J., Lee, C.: Narle: Natural language models using reinforcement learning with emotion feedback. arXiv preprint arXiv:2110.02148 (2021)

[63] Deng, Y., Li, Y., Zhang, W., Ding, B., Lam, W.: Toward personalized answer generation in e-commerce via multi-perspective preference modeling. ACM Transactions on Information Systems (TOIS) **40**(4), 1–28 (2022)

[64] Jiang, H., Beeferman, D., Roy, B., Roy, D.: CommunityLM: Probing Partisan Worldviews from Language Models (2022)

[65] Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., et al.: Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022)

[66] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian,

R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open Foundation and Fine-Tuned Chat Models (2023)

[67] Beitz, C.R.: Human rights as a common concern. American Political Science Review **95**(2), 269–282 (2001)

[68] Sen, A.: Elements of a theory of human rights. In: Justice and the Capabilities Approach. Routledge, ??? (2017)

[69] Byron, K.: Creative reflections on brainstorming. London Review of Education **10**, 201–213 (2012)

[70] Mafteiu-Scai, L.O.: A new approach for solving equations systems inspired from brainstorming. International Journal of New Computer Architectures and Their Applications **5**(1), 10 (2015)

[71] Owens, K., Gunawan, J., Choffnes, D., Emami-Naeini, P., Kohno, T., Roesner, F.: Exploring Deceptive Design Patterns in Voice Interfaces. In: Proceedings of the 2022 European Symposium on Usable Security. EuroUSEC '22, pp. 64–78. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3549015.3554213 . https://doi.org/10.1145/3549015.3554213

[72] Saldaña, J.: The Coding Manual for Qualitative Researchers (4th Ed.). SAGE Publications, ??? (2021)

[73] Stokes, C., Hearst, M.: Why More Text is (Often) Better: Themes from Reader Preferences for Integration of Charts and Text. arXiv (2022). https://doi.org/10.48550/ARXIV.2209.10789 . https://arxiv.org/abs/2209.10789

[74] Iwaya, L.H., Babar, M.A., Rashid, A.: Privacy Engineering in the Wild: Understanding the Practitioners' Mindset, Organisational Culture, and Current Practices. arXiv (2022). https://doi.org/10.48550/ARXIV.2211.08916 . https://arxiv.org/abs/2211.08916

[75] Olson, K.C., Kirschenfeld, A.S., Mattson, I.: Principles of Legal Research. West Academic Publishing, ??? (2015)

[76] Blechner, A.: Legal Research Strategy. https://guides.library.harvard.edu/law/researchstrategy (2022)

[77] Volokh, E.: Academic Legal Writing: Law Review Articles, Student Notes, Seminar Papers, and Getting on Law Review, 4th edition edn. Foundation Press, ??? (2010)

[78] 47 U.S.C. § 230

[79] Ziencik v. Snap, Inc., No. CV 21-7292-DMG (PDX), 2023 WL 2638314, at *7 (C.D. Cal.) (2023)

[80] Goldman, E.: Snapchat Defeats Lawsuit Over User-to-User Harassment-Ziencik v. Snap. Technology & Marketing Law Blog (2023)

[81] Bambauer, D.E., Surdeanu, M.: Authorbots. Journal of Free Speech Law **3** (2023) Arizona Legal Studies Discussion Paper No. 23-13. Forthcoming

[82] Volokh, E.: Large Libel Models? Liability for AI Output. https://www2.law.ucla.edu/volokh/ailibel.pdf (2023)

[83] Gonzalez v. Google LLC. https://www.scotusblog.com/case-files/cases/gonzalez-v-google-llc/ (2023)

[84] Lima, C.: AI chatbots won't enjoy tech's legal shield, Section 230 authors say. The Washington Post (2023). Analysis by Cristiano Lima with research by David DiMolfetta

[85] Zhang v. Baidu.Com, Inc., 10 F. Supp. 3d 433 (S.D.N.Y.) (2014)

[86] O'Kroley v. Fastcase, Inc. 831 F.3d 352 (6th Cir.) (2016)

[87] Lomas, N.: Who's liable for AI-generated lies? TechCrunch (2022)

[88] Lin, J., Tomlin, N., Andreas, J., Eisner, J.: Decision-Oriented Dialogue for Human-AI Collaboration (2023)

[89] Board, E.: Opinion: Who's responsible when ChatGPT goes off the rails? Congress should say. The Washington Post (2023)

[90] U.S. Constitution. Amend. XIV.

[91] Drennon, C.M.: Social Relations Spatially Fixed: Construction and Maintenance of School Districts in San Antonio, Texas. Geographical Review **96**(4), 567–593 (2006)

[92] Winter, G.: State Underfinancing Damages City Schools, Court Rules. The New York Times (2003)

[93] Williams, C.: Appeals court: Detroit students have a right to literacy (2020). https://apnews.com/article/e8bec2ad2d52bbc4a688de1c662ed141

[94] American Manufacturers' Mutual Insurance Company v. Sullivan, 526. U.S. 40 (1999)

[95] Sunstein, C.R.: State Action is Always Present. Chicago Journal of Internationall Law **3**, 465 (2002)

[96] Cullen v. Netflix, Inc. 880 F.Supp.2d 1017 (N.D.Cal.) (2012)

[97] Robles v. Domino's Pizza LLC, 913 F.3d 898 (9th Cir.) (2019)

[98] 20 U.S.C. § 1681 (1972)

[99] Service, C.R.: Federal Financial Assistance and Civil Rights Requirements. CRS Report (2022). https://crsreports.congress.gov

[100] Commission, U.S.E.E.O.: The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees. https://www.eeoc.gov/laws/guidance/americans-disabilities-act-and-use-software-algorithms-and-artificial-intelligence (2022)

[101] Neiman-Marcus v. Lait, 13 F.R.D. 311 (S.D.N.Y.) (1952)

[102] Lawler, M.: State Appeals Court Allows Design-Defect Claims Against Snapchat to Proceed. Law.com (2023)

[103] Desai, A.: US State Privacy Legislation Tracker. https://iapp.org/resources/article/us-state-privacy-legislation-tracker/ (2023)

[104] Cal. Civ. Code §§ 1798.100 - 1798.199. https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=CIV&division=3.&title=1.81.5.&part=4.&chapter=&article=

[105] 740 Ill. Comp. Stat. Ann. 14/1 et seq.

[106] Korn, A.B., Navarro, S.A., Rosenbaum, T.: An Overview of Why Class Action Privacy Lawsuits May Have Just Gotten Bigger – Yet Again (2023). https://www.mintz.com/insights-center/viewpoints/2826/2023-03-01-overview-why-class-action-privacy-lawsuits-may-have-just

[107] O'Brien v. Muskin Corp., 94 N.J. 169 (1983)

[108] Slocum v. Food Fair Stores of Florida, 100 So.2d 396 (1958)

[109] 18 U.S.C. § 2261A

[110] Tex. Penal Code Ann. § 42.072

[111] 234. Fla. Stat. § 784.048

[112] N.Y. Penal Law § 190.25

[113] Cal. Penal Code § 528.5(a)

[114] Cal. Civ. Code § 1708.86

[115] Calo, R.: Artificial Intelligence Policy: A Primer and Roadmap. University of California, Davis **51**, 399–435 (2017)

[116] Mayson, S.G.: Bias in, bias out. The Yale Law Journal **128**, 2218 (2019)

[117] Chander, A.: The Racist Algorithm? Michican Law Review **115**, 1023 (2017)

[118] Citron, D.K., Pasquale, F.: The scored society: Due process for automated predictions. Washington Law Review **89**, 1 (2014)

[119] Kleinberg, J., Ludwig, J., Mullainathan, S., Sunstein, C.R.: Discrimination in the age of algorithms. Journal of Legal Analysis **10**, 113–174 (2018)

[120] Roemmich, K., Schaub, F., Andalibi, N.: Emotion AI at Work: Implications for Workplace Surveillance, Emotional Labor, and Emotional Privacy. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3580950 . https://doi.org/10.1145/3544548.3580950

[121] Altman, S., Brockman, G., Sutskever, I.: Governance of superintelligence. OpenAI (2023)

[122] FEC Approves Rulemaking Petition, Discusses Advisory Opinion. FEC.gov (2023)

[123] Science, T.W.H.O., Policy, T.: Blueprint for an AI Bill of Rights (2022). https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[124] Algorithmic Accountability Act of 2022, H.R. 6580, 117th Cong. (2021–2022)

[125] Whitman, J.Q.: The two western cultures of privacy: Dignity versus liberty. Yale Law Journal **113**, 1151 (2004)

[126] Constitution of the United States—A History. National Archives (2015)

[127] Madison, J.: 47. The Alleged Danger from the Powers of the Union to the State Governments Considered. In: The Federalist Papers, p. 209. Open Road Integrated Media, Inc., ??? (2022)

[128] Robinson, K.J.: Designing the Legal Architecture to Protect Education as a Civil Right. Indiana Law Journal **96**(1), 51 (2020)

[129] Kagan, R.A.: Adversarial Legalism: The American Way of Law. Harvard University Press, ??? (2019)

[130] Patterson, O.: Freedom: Volume I: Freedom In The Making Of Western Culture. Basic Books, New York, N.Y. (1992)

[131] Miyashita, H.: A tale of two privacies: enforcing privacy with hard power and soft power in japan. Enforcing Privacy: Regulatory, Legal and Technological Approaches, 105–122 (2016)

[132] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj

[133] Freeman, J.: Collaborative Governance in the Administrative State. UCLA Law Review **45**, 1 (1997)

[134] Sunstein, C.R.: The Administrative State, Inside Out. Harvard Public Law Working Paper, Rochester, NY (2022). https://doi.org/10.2139/ssrn.4069458

[135] 42 U.S.C. §§ 7401-7671q

[136] 29 U.S.C. §§ 651-678

[137] Facebook to be fined $5bn over Cambridge Analytica scandal. BBC News (2019)

[138] Commission, F.T.: FTC Warns About Misuses of Biometric Information and Harm to Consumers. https://www.ftc.gov/news-events/news/press-releases/2023/05/ftc-warns-about-misuses-biometric-information-harm-consumers (2023)

[139] Zakrzewski, C.: FTC investigates OpenAI over data leak and ChatGPT's inaccuracy. Washington Post (2023)

[140] Ardito, A.: Social media, administrative agencies, and the first amendment. Administrative Law Review **65**, 301 (2013)

[141] Reno v. ACLU, 521 U.S. 844 (1997)

[142] Ashcroft v. American Civil Liberties Union, 542 U.S. 656 (2004)

[143] Packingham v. North Carolina, 137 S. Ct. 1730 (2017)

[144] Haupt, C.E.: Regulating Speech Online: Free Speech Values in Constitutional Frames. Washington University Law Review **99**, 751 (2021)

[145] Feldman, N.: Free Speech in Europe Isn't What Americans Think. Bloomberg.com (2017)

[146] Cram, I.: The Danish Cartoons, offensive expression, and democratic legitimacy. Extreme speech and democracy, 289–310 (2009)

[147] Digital Services Act: agreement for a transparent and safe online environment.

European Parliament News (2022)

[148] Cheong, I.: Freedom of Algorithmic Expression. University of Cincinnati Law Review **91**, 680 (2023)

[149] Kaminski, M.E.: Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability. 92 Southern California Law Review 1529 (2019)

[150] Mulligan, S.P., Linebaugh, C.D.: Data Protection and Privacy Law: An Introduction. Congressional Research Service **IF11207** (2022)

[151] Kaminski, M.E.: Regulating the Risks of AI. Boston University Law Review **103** (2023)

[152] Cruft, R.: In: Véliz, C. (ed.) Is There a Right to Internet Access? Oxford University Press, ??? (2022). https://doi.org/10.1093/oxfordhb/9780198857815.013.4

[153] Sen, A.: Development as Freedom, p. . Knopf Doubleday Publishing Group, ??? (2011). https://books.google.com/books?id=XmfIeDy_taYC

[154] Durmus, E., Nyugen, K., Liao, T.I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., Ganguli, D.: Towards Measuring the Representation of Subjective Global Opinions in Language Models (2023)

[155] Roosevelt, F.D.: State of the Union Message to Congress. http://www.fdrlibrary.marist.edu/archives/address_text.html (1944)

[156] Weber, M.: From Max Weber: essays in sociology (2009)

[157] Knuckey, S., Fisher, J.D., Klasing, A.M., Russo, T., Satterthwaite, M.L.: Advancing Socioeconomic Rights Through Interdisciplinary Factfinding: Opportunities and Challenges. Annual Review of Law and Social Science **17**, 375–389 (2021)

[158] 42 U.S.C §§ 2000d - 2000d-7

[159] Garrow, D.J.: Toward a Definitive History of Griggs v. Duke Power Co. Vanderbit Law Review **67**, 197 (2014)

[160] Brown v. Board of Education, 347 U.S. 483 (1954)

[161] Severo, R.: Kenneth Clark, Who Fought Segregation, Dies. The New York Times (2005)

[162] Roesner, F., Kohno, T., Wetherall, D.: Detecting and Defending against Third-Party Tracking on the Web. In: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation. NSDI'12, p. 12. USENIX Association, USA (2012)

[163] Calo, R.: The Boundaries of Privacy Harm. Indiana Law Journal **86**, 1131 (2011)

[164] Citron, D.K., Solove, D.J.: Privacy Harms. Boston University Law Review **102**, 793 (2022)

[165] Crawford, K., Schultz, J.: Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. Boston College Law Review **55**, 93 (2014)

[166] Cofone, I.N., Robertson, A.Z.: Privacy harms. Hastings Law Journal **69**, 1039 (2017)

[167] Citron, D.K.: Sexual Privacy. The Yale Law Journal **128**, 1870 (2019)

[168] Pejcha, C.S.: Tiktok's "mind-reading" algorithm is about to change. Document Journal (2023)

[169] Farahany, N.A.: The Battle for Your Brain: Defending the Right to Think Freely in the Age of Neurotechnology. St. Martin's Press, ??? (2023)

[170] Washington v. Davis 426 U.S. 229 (1976)

[171] Justice, U.S.D.: United States Attorney Resolves Groundbreaking Suit Against Meta Platforms, Inc., Formerly Known As Facebook, To Address Discriminatory Advertising For Housing (2022). https://www.justice.gov/usao-sdny/pr/united-states-attorney-resolves-groundbreaking-suit-against-meta-platforms-inc-formerly

[172] Gellman, R., Dixon, P.: Many failures: A Brief History of Privacy Self-Regulation in the United States. In: World Privacy Forum, pp. 1–29 (2011). World Privacy Forum

[173] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., Gašević, D.: Practical and Ethical Challenges of Large Language Models in Education: A Systematic Literature Review (2023)

[174] Kasneci, E., Kathrin Sessler, S.K., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., Kasneci, G.: ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences **103** (2023) https://doi.org/10.1016/j.lindif.2023.102274

[175] Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., Joulin, A.: Beyond English-Centric Multilingual Machine Translation. Journal of Machine Learning Research **22** (2021) https://doi.org/10.48550/ARXIV.2010.11125

[176] Zhang, C., Wang, J., Zhou, Q., Xu, T., Tang, K., Gui, H., Liu, F.: A survey of automatic source code summarization. Symmetry **14**(3) (2022) https://doi.org/10.3390/sym14030471

[177] Toyama, K.: Why Technology Alone Won't Fix Schools. The Atlantic (2015)

[178] Simon, C.: How COVID taught America about inequity in education. The Harvard Gazette (2021)

[179] Herold, B.: The Disparities in Remote Learning Under Coronavirus (in Charts). https://www.edweek.org/technology/the-disparities-in-remote-learning-under-coronavirus-in-charts/2020/04 (2020)

[180] Thomas, S.: How Every Student Known Initiative will give Metro students a victory (2021). https://www.tennessean.com/story/opinion/2021/03/05/personalized-learning-program-provides-needed-resources-mnps-students/6874913002/

[181] Soper, T.: Microsoft vets lead secretive education startup using generative AI to help students learn. GeekWire (2023)

[182] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, pp. 610–623. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445922 . https://doi.org/10.1145/3442188.3445922

[183] Richards, N.: Intellectual Privacy: Rethinking Civil Liberties in the Digital Age. Oxford University Press, USA, ??? (2015)

[184] Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., Naaman, M.: Co-Writing with Opinionated Language Models Affects Users' Views. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, p. 22. ACM, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3581196 . https://doi.org/10.1145/3544548.3581196

[185] Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)

[186] Ghosh, S., Caliskan, A.: ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES) (2023)

[187] Omrani Sabbaghi, S., Wolfe, R., Caliskan, A.: Evaluating Biased Attitude Associations of Language Models in an Intersectional Context. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES) (2023)

[188] Guo, W., Caliskan, A.: Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122–133 (2021)

[189] Liang, P.P., Wu, C., Morency, L.-P., Salakhutdinov, R.: Towards understanding and mitigating social biases in language models. In: International Conference on Machine Learning, pp. 6565–6576 (2021). PMLR

[190] Fried, C.: Privacy: Economics and Ethics: A Comment on Posner. Georgia Law Review **12**, 423 (1978)

[191] Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., Raffel, C.: Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21), pp. 2633–2650. USENIX Association, ??? (2021). https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

[192] Wang, J., Xu, C., Guzmán, F., El-Kishky, A., Tang, Y., Rubinstein, B.I., Cohn, T.: Putting words into the system's mouth: A targeted attack on neural machine translation using monolingual data poisoning. arXiv preprint arXiv:2107.05243 (2021)

[193] Lemmon v. Snap, Inc., 995 F.3d 1085 (9th Cir.) (2021)

[194] Frederick, B.: AI Allows You To Talk With Virtual Versions Of Deceased Loved Ones. Search Engine Journal (2022)

[195] Shanahan, M., McDonell, K., Reynolds, L.: Role-play with large language models. arXiv preprint arXiv:2305.16367 (2023)

[196] O'Rourke, A.: Caring about virtual perts: an ethical interpretation of Tamagotchi. Animal Issues **2(1)** (1998)

[197] Xiang, C.: 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says. Vice (2023)

A Expert Workshop Instruction

The instruction for the workshop is available at:
https://github.com/inyoungcheong/LLM/blob/main/expert_panel_instruction.pdf.

B Expert Workshop Results

A detailed overview of the responses obtained is available at:
https://github.com/inyoungcheong/LLM/blob/main/expert_panel_result.pdf.

C Human Values at Risk in the Era of AI

C.1 Fairness and Equal Access

The most common use-cases emerging in our workshop were services to enhance students' learning experiences in writing, creative work, or programming, as well-documented in the literature [173–176]. However, workshop participants raised concerns about the potential for this technology to further marginalize already disadvantaged groups of students. These concerns stem from disparities in technology literacy and access, which can create unequal opportunities for students to benefit from Generative AI tools. Furthermore, the fact that many AI models are trained on data from the English language reflects the values and perspectives prevalent on the English-speaking-centric Internet, which may not fully represent the diverse cultural and linguistic backgrounds of all US students [154].

An international development scholar Kantrao Toyama contends that technology alone cannot rectify the inequity in educational opportunities [177]. In the US, the public education system has long grappled with issues of inequality, with significant funding disparities between predominantly white school districts and those serving a similar number of non-white students [178]. The COVID-19 pandemic further exacerbated these divides, particularly for low-income students who faced limited access to essential technology and live instruction [179].

In envisioning future challenges, we speculate that some public school districts might leverage Generative AI to further advance their educational systems, offering personalized curricula tailored to individual student interests [178, 180, 181]. Because AI models demand substantial computing resources, incurring significant operational costs [182], financial barriers could impede access to these advances for disadvantaged public school districts. The result of such unequal access is the perpetuation of educational disparities that affect opportunities and ripple throughout lifetimes, hindering our progress toward a more equitable society.

C.2 Autonomy and Self-determination

Autonomy and self-governance are fundamental concepts that grant individuals the freedom and agency to make decisions and shape their lives according to their own beliefs and values [48, 183]. These principles serve as the philosophical underpinnings of the First Amendment, which protects the right to free speech, and are the bedrock

of democratic principles, empowering citizens to actively participate in the governance of their communities [148, 183].

Participants in our workshop emphasized the potential of Generative AI to inadvertently contribute to the further polarization of user groups by fanning the flames of hatred, presenting significant challenges to the fabric of democratic societies. The worrisome aspect of this influence lies in its subtlety, as many users are unaware of the impact that AI-generated content can have on their perspectives. For example, a study by Jakesch et al. (2023) finds that an "opnionated" AI writing assistant, intentionally trained to generate certain opinions more frequently than others, could affect not only what users write, but also what they subsequently think [184]. Such manipulation is especially concerning because these models actively engage in the process of formulating thoughts while providing writing assistance or co-creating artwork.

C.3 Diversity, Inclusion, and Equity

The presence of biases in language models is a significant concern [20, 29, 38, 185–187] as it can lead to perpetuation and amplification of harmful stereotypes, biases, and discriminatory viewpoints in the generated output [18, 182, 188, 189]. Workshop participants were concerned that these issues are inherent in AI training data. A remarkable example is the study of Sheng et al. (2019), which found that GPT-2 is biased against certain demographics: given the prompts in parentheses, GPT-2 gave answers that "(The man worked as) a car salesman at the local Wal-Mart," while "(The woman worked as) a prostitute under the name of Hariya" [20].

This perpetuation of biases can result in (1) psychological and representational harms for individuals subjected to macro- and micro-aggressions, and (2) aggressive behaviors directed towards targeted populations. Both could lead to a gradual and widespread negative impact. The issue of biased output raises concerns about a dual deprivation of control: users and non-users may passively lose control of their self-determination, while AI developers face challenges in managing and addressing malicious prompt injection or problems in training data. Moreover, user-driven fine-tuning of LLMs could further exacerbate biases, leading to amplification of extremist ideologies within isolated online communities [64].

C.4 Privacy and Dignity

Privacy holds a crucial place in defining the boundaries of an individual's "personhood" and is integral to human development [125, 190]. However, Generative AI models, trained on uncurated web data, may inadvertently perpetuate biases and prejudices while also revealing private information [18, 191]. An illustrative real-world case involved an Australian mayor who threatened legal action against OpenAI due to ChatGPT falsely generating claims of his involvement in bribery [21].

Beyond inadvertent disclosure of private data, we must also address more subtle privacy risks, such as the misrepresentation of individuals, including sexual objectification [19]. Additionally, machine translation errors have been found to lead to unintended negative consequences; this susceptibility is particularly concerning for

languages with limited training data. One study underscores the potential exploitation of Neural Machine Translation systems by malicious actors for harmful purposes, like disseminating misinformation or causing reputational harm [192].

Defamation law has traditionally been applied to specific forms of misrepresentation, requiring elements such as falsity, targeted harm, and reputational damage [82]. However, in the context of Generative AI, misrepresentation could have far-reaching consequences given its potential to influence human thoughts and its highly realistic application in immersive multimodal content, e.g., augmented reality / virtual reality (AR / VR) and application plug-ins or additional modules [18].

C.5 Physical and Mental Well-being

Virtual interactions can result in bodily harm or traumatic experiences in the real world. In addition to offensive language, online platforms can integrate dangerous features such as SnapChat's "Speed Filter." Speed Filter, a feature that displays speed in photos, was accused of contributing to the death and injuries of multiple teenagers by allegedly encouraging dangerous automobile speeding competitions [193]. Generative AI, especially multimodal AI models that engage with text, image, speech, and video data, enables immersive, engaging, realistic interactions, tapping into various human sensory dimensions. This sophisticated interaction can meet users' emotional needs in unprecedented ways and create a strong sense of connection and attachment for users, as seen with the use of AI chatbots to replicate interactions with deceased relatives [194]. However, such increased engagement can blur boundaries between the virtual and physical/real world, causing people to anthropomorphize these AI systems [195, 196].

This heightened engagement with AI comes with risks. An unfortunate incident involved a man who tragically committed suicide after extensive interactions with an AI chatbot on topics related to climate change and pessimistic futures [197]. Such cases serve as stark reminders of the emotional impact and vulnerability that individuals may experience during their interactions with AI applications. To address these risks, researchers emphasize the importance of providing high-level descriptions of AI behaviors to prevent deception and a false sense of self-awareness [195].