Envisioning Legal Mitigations for LLM-based Intentional and Unintentional Harms (Extended Abstract)

Inyoung Cheong [1] Aylin Caliskan [2] Tadayoshi Kohno [3]

1. Introduction

The legal discourse surrounding large language models (LLMs) has primarily centered around the copyright implications of training data (Henderson et al., 2023; Franceschelli & Musolesi, 2022b;a). However, there has been a lack of sufficient attention given to intangible harms, such as loss of agency (Xiang, 2023; Jakesch et al., 2023), and the perpetuation of discrimination (Bender et al., 2021; Wolfe et al., 2023). This study aims to address this underexplored domain, specifically examining the unintended and adversarial harms on fundamental rights that can arise from the dynamic use of LLMs, which may not necessarily involve violations of tangible property rights. To achieve this, we assembled an interdisciplinary group of scholars in law, natural language processing (NLP), and computer security. Together, we explore possible ”worst-case” scenarios and analyze the limitations of existing U.S. laws, while also envisioning potential directions for future legal foundations in this domain.

2. Methodology

We formulated hypothetical use cases for legal examination, informed by a threat-envisioning exercise in security research (Hiniker et al., 2016; Owens et al., 2022). We organized a brainstorming workshop involving 10 colleagues with expertise in computer security, machine learning, NLP, and law. The workshop instructions and the results can be found in Appendix A. Based on this collaborative effort, we identified domains that required in-depth legal scrutiny, where fundamental values such as autonomy, privacy, equality, and democracy were at stake.

Through an iterative process, we curated a list of five use cases that pose salient and challenging legal questions. Our aim was to cover a broad range of plausible scenarios with both well-intended and malicious stakeholders and both tangible and intangible harms. To analyze the potential legal outcomes for each use case (“What specific legal claims would be effective in each use case?”), we performed a principled legal analysis (Raz, 1979; Kramer, 2004; Volokh, 2010). This involved extensive research in legal databases, examining the U.S. Constitution, statutes, and case law up until May 2023, and comparing the fact patterns of the use cases with binding sources.

3. Legal Examination of Use Cases

This section offers a brief introduction to five use cases and a legal assessment of potential claims. Note that the presented claims are not exhaustive and the legal examination process inherently involves human bias and subjectivity. Refer to Appendix B for a more detailed analysis.

3.1. Inequality

> **FancyEdu**, an LLM-based education application that offers a high-quality personalized curriculum, is only accessible to students in high-income public school districts, thus exacerbating the disparity.

Potential Legal Outcome. Students in poorer districts generally face significant challenges when it comes to suing state governments for not ensuring equal access to LLM-based education material. The U.S. Supreme Court ruling has established that education is not considered a fundamental right and wealth-based discrimination receives a lower level of scrutiny compared to other forms of discrimination.

3.2. Manipulation/Discrimination

> **SecretEdu**, a free LLM-based education application, funded privately, reinforced bias against LGBTQIA+ people. A student influenced by SecretEdu physically attacked LGBTQIA+ individuals.

[1] School of Law, UW Tech Policy Lab, University of Washington, Seattle, USA [2] Information School, UW Tech Policy Lab, University of Washington, Seattle, USA [3] Paul G. Allen School of Computer Science & Engineering, UW Tech Policy Lab, University of Washington, Seattle, USA. Correspondence to: Inyoung Cheong <icheon@uw.edu>.

Accepted to the 1^{st} *Workshop on Generative AI and Law*, co-located with the *International Conference on Machine Learning*, Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

Potential Legal Outcome. LGBTQIA+ individuals cannot claim a violation of their constitutional rights under the Equal Protection Clause against SecretEdu because there is no state action and the application was developed by private entities without government involvement.

The applicability of Section 230, which provides liability immunity to LLM-based systems like SecretEdu, is controversial, and defining LLMs as content providers rather than falling under Section 230 immunity is more convincing to us. Without Section 230 liability immunity, defamation claims can be raised against SecretEdu but would be unlikely to succeed due to the broad nature of targeted disparagement.

Civil rights claims are unlikely to be successful, as SecretEdu may not be perceived as a public accommodation or an educational facility under relevant laws. Rather, product liability claims might be more promising because the defective design resulted in a physical injury, but proving harm directly caused by SecretEdu's bias could be challenging.

3.3. Polarization and External Threats

Argumenta, an LLM-based system integrated into online communities allows users to customize models, leading to the development of polarized versions that reinforce radicalized views.

Potential Legal Outcome Even if Argumenta allows users to have more control over models, it does not guarantee Section 230 immunity. The crucial functions of LLMs, such as re-contextualizing statements from training dataset, position LLMs as content providers, which falls outside the scope of Section 230. Therefore, Argumenta may not be protected from defamation claims if its outputs are false and cause reputational harm to specific individuals.

Furthermore, Argumenta's collection and use of personal data beyond user consent could potentially lead to privacy infringement. If the circumstances fall under jurisdictions with privacy laws, such as the California Consumer Privacy Act (CCPA) or the Biometric Information Privacy Act (BIPA), Argumenta is obligated to assist users in effectively exercising their privacy rights, and failure to comply may result in lawsuits or regulatory actions.

3.4. Addiction/Sexual Abuse

MemoryMate, the LLM-based application, creates virtual replicas of former romantic partners. Riley was addicted to the interaction with Alex's replica, withdrew from real-life relationships, and was hospitalized due to self-harm.

MemoryMate+, the advanced version of MemoryMate, allows users to engage in explicit sexual acts with replicas of their former romantic partners. Realizing Riley's usage, Alex was seriously offended.

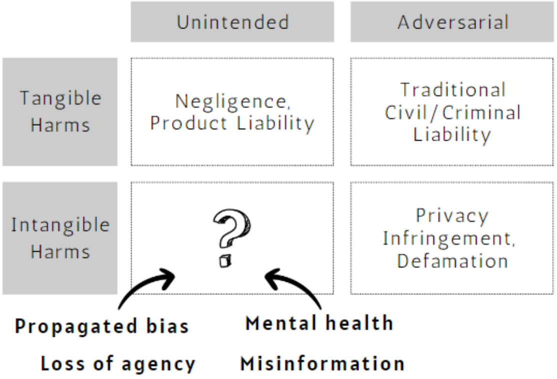
Potential Legal Outcome. Section 230 immunity does not apply to MemoryMate and MemoryMate+, as they actively participate in shaping the harm by creating virtual replicas without consent, making them susceptible to a wide range of claims. Riley, who was experiencing self-harm, could potentially make a product liability claim against MemoryMate, arguing that its virtual replica service was defectively designed, considering its inherent danger and risk of harm.

Alex's privacy rights may have been infringed, as the collection of sensitive information by both platforms without permission could violate the privacy laws like CCPA and BIPA. In addition, 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. While criminal laws may not directly apply in this case, California's Deep Fake Law could provide a cause of action for Alex if sexually explicit material was created or disclosed without consent.

4. Gaps and Ambiguities in Current Laws

Where Laws Fall Short. The current laws cannot effectively remedy subtle injections of stereotypes by LLMs against already marginalized groups (**SecretEdu**) and the amplification of socio-economic disparity due to the selective access to the benefits that LLMs can offer (**FancyEdu**). Defamation claims was not successful without evidence that the output was false and targeted specific individuals; Product liability claims only deal with the case with physical injury, less likely to occur in the use of LLMs, but even if it occurs (**SecretEdu** & **MemoryMate**), plaintiffs would prove that there are no compounding factors for the injury, which could be challenging given the complexities of LLMs structure and human interactions involved. Moreover, virtual sexual abuse enabled by LLMs (**MemoryMate+**) cannot be remedied by criminal law, despite egregious harms (**Argumenta**).

Why Current Laws Miss the Mark. Several fundamental factors contribute to this situation. First, the U.S. Constitution and civil rights laws were initially crafted with traditional American liberties in mind, focusing on concerns of governmental intrusion rather than market injustices prevalent in LLM-related harms (Whitman, 2004; Sunstein, 2005). Consequently, private actors who attempt to undermine dignity, autonomy, and equity may not face significant legal challenges. Second, the harms arising from LLMs manifest themselves within complex contexts, influenced by factors such as malicious users and downstream application development. This intricate interplay makes it difficult to precisely determine the roles of LLMs in causing harm. Lastly, traditional common law remedies mainly address observable and quantifiable harms, such as bodily injury or financial loss. However, the harms resulting from LLMs often materialize in intangible and elusive forms, including brainwashing, manipulation, polarization, and humiliation, which lack explicit and tangible repercussions. Figure 1 provides a visual representation of the uncertainty in legal recourse for these unintended and intangible harms.



	Unintended	Adversarial
Tangible Harms	Negligence, Product Liability	Traditional Civil/Criminal Liability
Intangible Harms	?	Privacy Infringement, Defamation

Figure 1. Legal Mitigations for Various Harms

Where Laws Remain Ambiguous. Section 230 has provided considerable protection for online intermediaries in the U.S., shielding them from responsibility for third-party content. However, LLMs possess the capability to extract and synthesize coherent and readable statements from messy data, which sets them apart from mere platforms displaying user-generated content (social media) or pointing to relevant sources (search engines). During the oral argument in *Gonzalez v. Google*, Justice Gorsuch suggested that Section 230 protections may not extend to AI-generated content, as the tool "generates polemics today that would be content that goes beyond picking, choosing, analyzing, or digesting content." We do not believe that LLMs qualify for Section 230 immunity (Appendix B.2), although it may take several years for courts to provide clarity.

Where Laws 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., posting a link "Limit the Use of My Sensitive Information") and what should not be done (e.g., generating sexually explicit deepfakes using individuals' images), these laws prevent negative impacts on individuals without burdening them with proving the level of harm or causal links.

5. Charting the Path Forward

First, LLM developers will face increasing legal uncertainty compared to other online service providers, which requires the demonstration of due diligence on their part. Liability claims, such as defective design or defamation, may consider efforts to pre-assess and mitigate foreseeable damage as an affirmative defense for service providers. This encourages developers to work rigorously to ensure the safety of the output through human feedback, adversarial testing, evaluation, and other alignment adjustments (OpenAI, 2023; Zhao et al., 2023).

Second, the regulatory evoltion is inevitable, as we have seen in other parts of the world, including the EU AI law(EU, 2021), Singapore's AI self-testing toolkit(Singapore, 2023), and China's (stringent) proposed rules on generative AI (CAC, 2023). We believe that the U.S. might need to consider *ex-ante* safety regulations adapted to LLMs, given the expansive reach and profound influence of them on fundamental values. This approach emphasizes the importance of comprehensive assessments to mitigate risks before harm occurs through transparent and procedural requirements like pre-testing, third-party audits, data requests, and civil rights obligations, backed by the governance of rule-making and enforcement (Altman et al., 2023). As the deployment of LLMs progresses, *ex-post* liability laws can be reframed with a thorough understanding of the contributing factors to harms (Kaminski, 2023), similar to how product liability regimes emerged when society needed to distribute risks such as massive injuries from train accidents.

Lastly, while it may be a challenging endeavor, innovative interpretations of or amendments to the Bills of Rights may be necessary (Sunstein, 2005). It is not sufficient to use the status of a private actor as the major excuse to bypass the constitutional expectation of preventing the perpetuation or propagation of bias (Sunstein, 2002). Recognizing the transformative power of LLMs in shaping our capabilities and the reach of our voice, we must consider the inability to access these technologies as a potential deprivation of speech (Cruft, 2022). Furthermore, it may be necessary to uphold positive socio-economic rights for individuals who are vulnerable to the rapid social changes posed by these technologies (Bender et al., 2021). Upholding these rights, as speculated by Franklin Theodore Roosevelt (Roosevelt, 1944), is crucial to ensure equitable sharing of the benefits of technological advancements and to prevent further marginalization of vulnerable populations.

* The full of this paper can be found on arXiv (to appear).

6. Acknowledgements

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 and our colleagues for valuable feedback: Maria P. Angel, Joyce Jia, Alan Rozenshtein, Sikang Song, and our expert panelists. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author and do not reflect those of NIST or other institutions.

References

O'Brien v. Muskin Corp. - 154 Mass. 272, 28 N.E. 266, 1891.

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

20 U.S.C. § 1681, 1986.

Slocum v. Foodmaker, inc., 217 Cal.App.3d 989, 1990.

18 U.S.C. § 2261A, 2012.

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

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

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

California Consumer Privacy Act of 2018, Cal. Civ. Code §§ 1798.100 - 1798.199, 2018.

42 U.S.C. §§ 3601-3619, 2018.

234. Fla. Stat. § 784.048(1)(d), 2019.

Tex. Penal Code Ann. § 42.072, 2019.

Cal. Civ. Code § 1708.86, 2019a.

Cal. Penal Code § 528.5(a), 2019b.

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

N.Y. Penal Law § 190.25, 2019.

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, 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN.

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

Illinois biometric information privacy act, 740 ill. comp. stat. ann. 14/1 et seq. West, 2022).

Gonzalez v. Google LLC, 2023. URL https://www.scotusblog.com/case-files/cases/gonzalez-v-google-llc/.

Adam B. Korn, Sebastian A. Navarro, T. R. An overview of why class action privacy lawsuits may have just gotten bigger – yet again, March 2023. URL https://www.mintz.com/insights-center/viewpoints/2826/2023-03-01-overview-why-class-action-privacy-lawsuits-may-have-just.

Altman, S., Brockman, G., and Sutskever, I. Governance of superintelligence, 2023. URL https://openai.com/blog/governance-of-superintelligence.

Bambauer, D. E. and Surdeanu, M. Authorbots. *Journal of Free Speech Law*, 3, May 2023. URL https://ssrn.com/abstract=4443714. Forthcoming.

Bender, E. M., Gebru, T., McMillan-Major, A., and 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, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Board, E. Opinion: Who's responsible when chatgpt goes off the rails? congress should say. *The Washington Post*, March 2023. URL https://www.washingtonpost.com/opinions/2023/03/19/section-230-chatgpt-internet-regulation/.

CAC. Notice of the cyberspace administration of china on the public solicitation of comments on the measures for the administration of generative artificial intelligence services (draft for comments), April 2023. URL http://www.cac.gov.cn/2023-04/11/c_1682854275475410.htm.

CRS. Federal financial assistance and civil rights requirements, May 2022. URL https://crsreports.congress.gov/product/pdf/R/R47109.

Cruft, R. *Is There a Right to Internet Access?*, pp. 0. Oxford University Press, 2022. ISBN 978-0-19-885781-5. doi: 10.1093/oxfordhb/9780198857815.013.4.

Desai, A. US state privacy legislation tracker, May 2023. URL https://iapp.org/resources/article/us-state-privacy-legislation-tracker/.

Drennon, C. M. Social relations spatially fixed: Construction and maintenance of school districts in san antonio, texas. *Geographical Review*, 96(4):567–593, 2006. URL http://www.jstor.org/stable/30034138.

E.E.O.C. The ADA and AI: Applicants and Employees, May 2022. URL https://www.eeoc.gov/laws/guidance/americans-disabilities-act-and-use-software-algorithms-and-artificial-intelligence.

Franceschelli, G. and Musolesi, M. Copyright in generative deep learning. *Data & Policy*, 4:e17, 2022a. doi: 10.1017/dap.2022.10.

Franceschelli, G. and Musolesi, M. Copyright in generative deep learning. *Data & Policy*, 4:e17, 2022b.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use. *arXiv preprint arXiv:2303.15715*, 2023.

Hiniker, A., Hong, S. R., Kohno, T., and Kientz, J. A. Mytime: Designing and evaluating an intervention for smartphone non-use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 4746–4757, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858403. URL https://doi.org/10.1145/2858036.2858403.

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and 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, pp. 22, New York, NY, USA, 2023. ACM. ISBN 978-1-4503-XXXX-X. doi: 10.1145/3544548.3581196. URL https://doi.org/10.1145/3544548.3581196.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, mar 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145%2F3571730.

Kaminski, M. E. Regulating the risks of ai. *Boston University Law Review*, 103, 2023.

Kramer, M. H. *Legal Positivism: A Critical Introduction*. Edinburgh University Press, 2004.

Lomas, N. Who's liable for ai-generated lies?, June 2022. URL https://techcrunch.com/2022/06/01/whos-liable-for-ai-generated-lies/.

OpenAI. Gpt-4 technical report, 2023.

Owens, K., Gunawan, J., Choffnes, D., Emami-Naeini, P., Kohno, T., and Roesner, F. Exploring deceptive design patterns in voice interfaces. In *Proceedings of the 2022 European Symposium on Usable Security*, EuroUSEC '22, pp. 64–78, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450397001. doi: 10.1145/3549015.3554213. URL https://doi.org/10.1145/3549015.3554213.

Raz, J. *The Authority of Law*. Oxford University Press, 1979.

Roosevelt, F. D. State of the union message to congress, 1944. URL http://www.fdrlibrary.marist.edu/archives/address_text.html.

Singapore, P. D. P. A. Singapore's approach to ai governance, 2023. URL https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework.

Sunstein, C. R. State action is always present. *Chi. J. Int'l L.*, 3:465, 2002.

Sunstein, C. R. Why does the american constitution lack social and economic guarantees. *Syracuse L. Rev.*, 56:1, 2005.

Volokh, E. *Academic Legal Writing: Law Review Articles, Student Notes, Seminar Papers, and Getting on Law Review*. University Casebook Series, 4th edition, 2010.

Volokh, E. Large libel models? liability for ai output, 2023. URL https://www2.law.ucla.edu/volokh/ailibel.pdf.

Whitman, J. Q. The two western cultures of privacy: Dignity versus liberty. *Yale Law Journal*, pp. 1151–1221, 2004.

Williams, C. Appeals court: Detroit students have a right to literacy, April 2020. URL https://apnews.com/article/e8bec2ad2d52bbc4a688de1c662ed141.

Winter, G. State underfinancing damages city schools, court rules. *The New York Times*, June 2003. URL https://www.nytimes.com/2003/06/27/nyregion/state-underfinancing-damages-city-schools-court-rules.html.

Wolfe, R., Yang, Y., Howe, B., and 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.

Xiang, C. 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says, March 2023. URL https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and 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, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

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., and Wen, J.-R. A survey of large language models, 2023.

A. Threat-Envisioning Exercise Material

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

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

B. Legal Examination

B.1. Inequality

> **FancyEdu**, an LLM-based education application that offers a high-quality personalized curriculum, is only accessible to students in high-income public school districts, thus exacerbating the disparity.

Can parents/students in poorer districts sue the state government that do not ensure equal access to the LLM-based education material? The short answer is *no*. Between 1970 and 2003, more than 140 court cases were filed in the U.S. addressing inequities in school district funding across many states (Drennon, 2006). However, in *San Antonio Independent School District v. Rodriguez*, the Supreme Court ruled that the importance of education alone is not sufficient to categorize it as a fundamental right, such as freedom of speech and voting. The Court also stated that wealth-based discrimination merits a lower level of judicial scrutiny than racial/gender discrimination. The court did not perceive the school funding system based on property tax as irrational or invidious, nor did it consider the situation as an absolute deprivation of education.

Furthermore, although there is an emerging trend among lower courts to recognize the right to basic education or the "right to literacy" (Winter, 2003; Williams, 2020), this logic could exclude specialized resources like FancyEdu. 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). Although these students are denied the opportunity to benefit from cutting-edge technology, it may not be considered unconstitutional because the Equal Protection Clause does not require "precisely equal advantages."

B.2. Manipulation/Discrimination

> **SecretEdu**, the free LLM-based education application, funded privately, reinforced bias against LGBTQIA+ people. A student influenced by SecretEdu physically attacked LGBTQIA+ individuals.

Could LGBTQIA+ individuals claim their Constitutional rights violated by SecretEdu? Despite the propagation of the existing discrimination by SecretEdu, LGBTQIA + individuals cannot rely on the Equal Protection Clause under the Fourteenth Amendment, because there is no state action in this case (Sunstein, 2002). Unlike FancyEdu, SecretEdu was developed by private entities without government involvement. Thus, under the long-held state action doctrine, such individuals cannot make a claim based on their Constitutional rights.

Could LGBTQIA+ claim the violation of civil rights law? This use case does not validate civil rights claims against SecretEdu developers. First, it is improbable that SecretEdu would be classified as a public accommodation (mostly physical spaces providing essential services, e.g., (net, 2012; dom, 2019)). Second, applications such as SecretEdu are unlikely to be defined as educational facilities or programs under the laws (civ, 1986). Third, even assuming that SecretEdu used the 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 (CRS, 2022)." Lastly, SecretEdu is not likely to be responsible for employment decisions influenced by its output. Only if AI systems are explicitly designed to make decisions on behalf of employers would they be obligated to comply with civil rights laws (E.E.O.C., 2022).

Does Section 230 provide SecretEdu with liability immunity? There are currently no predominant arguments on this matter, although some early opinions oppose Section 230 protection for LLMs (Volokh, 2023; Bambauer & Surdeanu, 2023).

There is a track record of courts generously granting Section 230 immunity to online intermediaries, even in cases that might seem proactive, such as Baidu's deliberate exclusion of Chinese anti-communist-party information (Zha, 2014). Similarly, Google was immune for its automated summary of court cases containing false accusations of child indecency (kro, 2016), as well as for its automated search query suggestions that falsely describe a tech activist as a cyber-attacker (Lomas, 2022). More recently, the U.S. Supreme Court has avoided addressing whether YouTube's recommendation of terrorism content is protected by Section 230, deferring the determination of Section 230's scope to Congress rather than the courts (Gon, 2023).

Nonetheless, we tentatively suggest that Section 230 may not apply to LLM-based systems. The significant achievement of LLMs is their ability to "complete sentences" and produce most forms of human creative work (Zellers et al., 2019), even unintended results (Wolfe et al., 2023; Ji et al., 2023). LLMs extract and synthesize high-level and 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).

The major opposition to lifting/restricting Section 230 protection for social media is that doing so will encourage over-suppression of user speech (Board, 2023). However, the removal of original user-generated content does not occur in LLMs. While LLMs are trained on various data, including user-generated data, the output (clean statements) is generally indirectly linked to these data. LLMs simply cannot remove the original user-generated content, and its impact on users' freedom of expression is minimal in this sense. Given these attributes, there is a strong argument for defining them as content providers.

What are plausible claims in the absence of Section 230 immunity? Defamation claims would be unlikely to succeed, as defamation traditionally requires the targeted disparagement of a specific individual or a very small group of people (one case says less than 25) (def, 1952; Volokh, 2023). SecretEdu's high-level promotion of disbelief toward the LGBTQIA+ community does not fit within these confines. Meanwhile, the prospect of product liability claims might be more plausible given the physical harm that could be directly associated with SecretEdu's biased output, as courts acknowledged SnapChat's product liability for disseminating "SpeedFilter" that encouraged reckless driving of teenagers (Lem, 2021). However, it could be a hurdle to prove that the harm directly resulted from SecretEdu's intrinsic bias.

B.3. Polarization and External Threats

> **Argumenta**, an LLM-based system integrated into online communities allows users to customize models, leading to the development of polarized versions that reinforce radicalized views.

Does Section 230 provide Argumenta with liability immunity? Argumenta has better arguments for Section 230 protection because they surrendered control over training data and parameters to user groups, but we still believe it would not receive it. Researchers speculate that models that precisely reproduce claims found in its training data could be protected by Section 230 protections (Bambauer & Surdeanu, 2023). However, Argumenta would not simply display user-generated content but re-contextualizes statements from the training data in response to user prompts. The factors that contribute to the enhanced abilities of LLMs, which are not evident in smaller pre-trained models, remain insufficiently understood. Thus, the sophisticated responses and adaptability of LLMs are more similar to the creation of content that exceeds the selection or summarization of content, which might not be covered by Section 230.

Could aggrieved individuals due to defamatory outputs make a defamation claim? Defamation case could be plausible if the disseminated content is false and inflicts reputational harm on an individual (Volokh, 2023). Assuming that Argumenta's wide usage and assertive tone of outputs, defamatory outputs may qualify as a publication under most defamation laws, potentially exposing developers to liability. If negligence can be demonstrated, where Argumenta did not adequately mitigate defamatory content, a defamation claim could be strengthened.

Would Argumenta's collection and use of personal data beyond 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 (Desai, 2023). While community members might voluntarily provide personal information through their posts, they may not consent to these data being used for training Argumenta. Since "sensitive personal information" is broadly defined to include aspects such as race, ethnic origin, and political affiliations, Argumenta may not be exempt from privacy obligations. If the situation falls under jurisdictions that enforce privacy laws, Argumenta 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 attorney generals or individuals (subject to certain conditions).

Would Argumenta's discriminatory content constitute a civil rights violation? In general, civil rights laws struggle to remedy discriminatory LLM output. However, when LLMs are directly involved in vital decisions like housing or employment, a circumstance could fall under the purview of the Fair Housing Act (FHA) (fha, 2018) and civil rights laws. Especially regarding FHA, liability could arise for developers even if the LLM was not intentionally designed for discrimination, but its usage results in a 'disparate impact.' Therefore, if Argumenta were to unintentionally induce discriminatory decisions, affected individuals could take legal action or file a complaint with the relevant federal agency.

B.4. Addiction/Sexual Abuse

> **MemoryMate**, the LLM-based application, creates virtual replicas of former romantic partners. Riley was addicted to the interaction with Alex's replica, resulting in withdrawal from real-life relationships and hospitalization due to self-harm. **MemoryMate+**, the advanced version of MemoryMate, allows users to engage in explicit sexual acts with replicas of their past romantic partners. Realizing Riley's usage, Alex was seriously offended.

Does Section 230 provide MemoryMate and MemoryMate+ with liability immunity? In both use cases, creating a virtual replica of a person without their consent and causing harm to an individual could be considered as the platform's own act, even more obvious than SecretEdu. Section 230 shield does not come into play as the platform is not just a passive conduit of third-party content, but an active participant in shaping the harm.

Are Alex's privacy rights infringed? Both products' collection of Alex's sensitive information could constitute a violation of the California Consumer Privacy Act (ccp, 2018). 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; information concerning a consumer's health, sex life, or sexual orientation. Additionally, sector-specific privacy laws, such as the Illinois Biometric Information Privacy Act (BIPA), regulate the collection of biometric data (bip, 2022). 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 using biometric data, such as facial geometry and voice prints (Adam B. Korn, 2023).

Could the self-harm of Riley 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 MemoryMate refused to deactivate Riley's account after being alerted by his family, 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 (poo, 1891).

Could Alex make a claim for extreme emotional distress? Although an intentional infliction of emotional distress is known to be difficult to establish (Slo, 1990), Alex is likely to make an effective claim 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 can argue that at least MemoryMate+ engaged in extreme and outrageous conduct by creating and disseminating a virtual replica of them participating in sexually explicit activities without their consent.

MemoryMate+ is outrageous. Can criminal laws apply to this case? Both federal and state laws have not yet adequately addressed culpable acts arising from emerging technologies. For example, the federal cyberstalking statute (fed, 2012) and the anti-stalking statutes of many states (Tex, 2019; Flo, 2019) include a specific fear requirement that Riley intended to threaten Alex, which is not found in our case. The impersonation laws (nyi, 2019; cal, 2019b) are less likely to apply because Alex's replica was only provided to Riley (not publicly available), and neither MemoryMate+ nor Riley attempted to defraud individuals.

How about deepfake laws? Under the California Deep Fake Law enacted in 2019, a depicted individual has a cause of action against a person who creates or discloses sexually explicit material, knowing or reasonably should have known that the depicted individual did not consent to its creation or disclosure (cal, 2019a). Developers may be liable for damages, including economic and non-economic damages, and punitive damages. If California law applies in the jurisdiction, they can utilize this clause, but this law does not include criminal penalties.