The overarching aim of my research is to position law and policy as the key mechanisms through which we can achieve an AI ecosystem aligned with human values. As algorithmic systems proliferate across domains, from creative pursuits to mental health counseling to employment decisions, people are delegating more tasks and decisions to opaque AI systems due to their convenience and perceived benefits. However, the remarkable capabilities of these systems, combined with their lack of interpretability, create a *double vulnerability*—users come to heavily rely on them yet remain unable to assess their integrity. My research seeks to address this power imbalance by developing a legal and policy framework centered on ethics and accountability.

Through interdisciplinary collaboration with experts in law, computer security, and human-computer interaction (HCI), my scholarship addresses complex legal and ethical issues in AI governance frameworks and suggest adaptive mechanisms upholding dignity, equity, and justice. Towards the objective of realizing ethics-informed technology, my research outputs provide not only rigorous legal analysis, but also through direct techniques to embed communal values and judgments into AI systems. I plan to continue efforts in three key directions:

- **Legal Mitigations for AI-Mediated Harms**—conducting scenario analysis to reveal emerging threats posed by AI systems to legal rights and human values. Identifying regulatory gaps and constructing adaptive legal frameworks to uphold autonomy, equity and justice.
- **Case-based Reasoning for AI Alignment**—developing participatory methods to engage experts and publics in constructing evolving case repositories. Encoding pluralistic communal norms and precedents to inform ethical AI alignment grounded in context.
- **Rethinking Concepts of Individual Liberty**—showing the murkiness of applying traditional legal concepts (e.g. *freedom of expression, privacy*) to emerging technologies. Projecting new concepts of individual rights and responsibilities tailored to human-AI interactions.

1 Legal Mitigations for AI-Mediated Harms

The rise of AI chatbots has ignited both excitement and trepidation, with some fearing a potential takeover of human roles by machines [2]. To rationally address these concerns, my team conducted a threat-envisioning exercise with experts from machine learning, natural language processing (NLP), computer security, and law and policy [8, 7]. During the workshop, participants identified: **potential use-cases**, **stakeholders** affected by the systems, **datasets** used for the development of the systems, and **expected impacts** on stakeholders or society as a whole. Based on the findings during the workshop, we constructed five hypothetical cases that (1) portray both **beneficial** and **detrimental** applications of AI systems; (2) involve both **intentional** and **unintentional** harm by AI companies; and (3) entail **real-world harms** as well as **intangible virtual harms**.

By applying existing legal remedies to these hypothetical cases, we conclude that traditional approaches within US legal systems, whether 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 (Figure 1). A patchwork of domain-specific laws and the case-law approach fail to establish comprehensive risk management strategies that extend beyond isolated instances. Given the necessity of democratic oversight, the structural and unpredictable risks of AI systems, and the users' vulnerability to opaque systems, we propose to steer our regulatory system to govern AI by encoding human values into law as depicted in Figure 2.

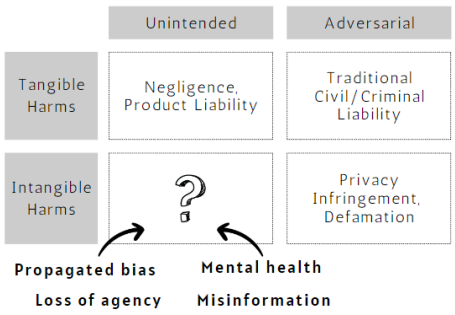

Figure 1: Legal Mitigations for AI-Mediated Harms.

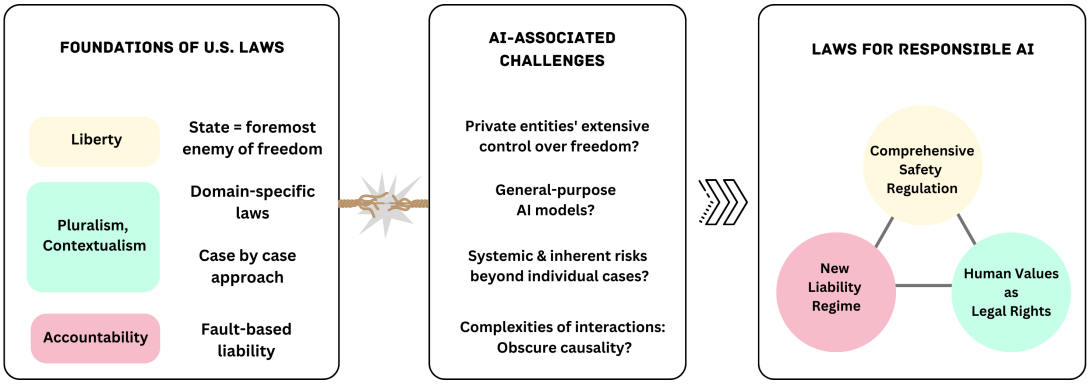

Figure 2: Responsible AI Regulatory Framework.

Future Work: What Types of AI Safety Governance? Many different proposals have been discussed that put guardrails on AI systems. They include the creation of a federal agency, the licensing of high-risk AI systems, third-party auditing, transparency reporting, mandates on safety and alignment, and other measures. I aim to investigate the unique nature of general-purpose AI models as compared to other technologies like cloud computing or search engines and propose pragmatic governance measures matched to the specific characteristics and rapid iterative development processes of foundation models. More specifically, I want to evaluate a myriad of regulatory techniques, including: (1) individual users' class actions versus government agencies enforcing rules; (2) Congress enacting legislation versus standard-setting organizations crafting voluntary best practices; (3) imposing substantive requirements (e.g. banning copyright-infringing training data) versus procedural requirements (e.g. mandatory opt-out processes); and (4) differentiated requirements based on whether systems are open-source versus proprietary.

2 Case-based Reasoning for AI Alignment

Complexities and ambiguities arise when we consider how AI should be aligned in practice: when faced with vast quantities of diverse (or conflicting) values from different individuals and communities, with *whose* values is AI to align, and *how* should AI do so? Supported by OpenAI [17, 10], my team proposes a complementary approach to rule-based AI alignment (e.g. [1]), grounded in ideas from the case-based reasoning (CBR) on the construction of policies through judgments on a set of cases [12].[1] Such a *case repository* could assist in AI alignment, both through acting as precedents to ground acceptable behaviors, and as a medium for individuals and communities to engage in moral reasoning around AI.

We present a human-AI hybrid process for creating a rich set of cases to support discussion around ethically-informed AI systems. The key steps are: (1) Collect a small set of **seed cases** from online communities and existing case studies/case law; (2) Recruit **domain experts** and facilitate workshops where experts identify key dimensions that impact the appropriateness of different AI responses, then use LLM to generate **new synthetic cases** along the key dimensions identified by experts; and (3) Engage **crowd workers** to evaluate the appropriateness of different AI responses to cases and refine synthetic cases to improve quality.

Lawyers Weighing in AI's Legal Advice As a part of the case-based deliberation project, I led research on examining guiding principles on AI chatbots providing legal guidance to the public [9]. We contribute a structured expert analysis with 20 legal professionals to uncover nuanced policy considerations around using LLMs for professional advice, using methods inspired by case-based reasoning. We convened workshops with 20 legal experts and elicited dimensions on appropriate AI assistance for sample user queries ("cases"). We categorized our expert dimensions into: (1) **user attributes**, (2) **query characteristics**, (3) **AI capabilities**, and (4) **impacts**. Beyond known issues like hallucinations, experts revealed novel legal problems, including that users' conversations with LLMs are not protected by attorney-client confidentiality or bound to professional ethics that

[1] https://social.cs.washington.edu/case-law-ai-policy/.

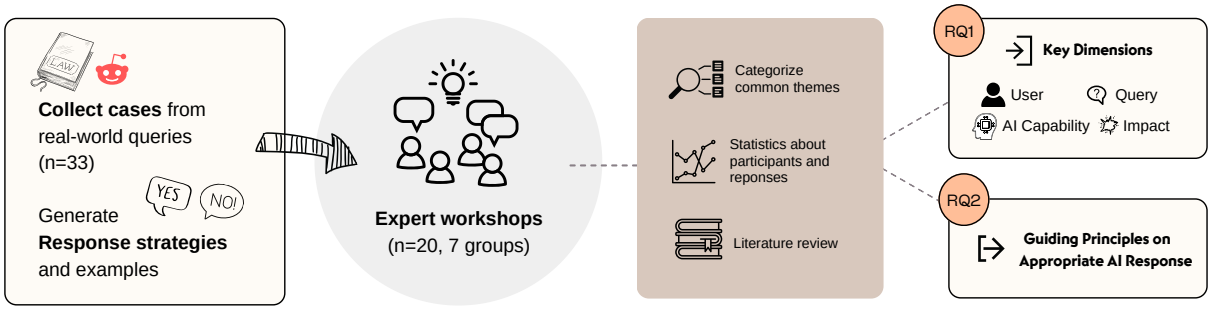

Figure 3: Overview of our research process and findings.

guard against conflicted counsel or poor quality advice. This accountability deficit led participants to advocate for AI systems to help users polish their legal questions and relevant facts, rather than recommend specific actions.

Throughout this process, we found long-standing legal scholarship on the unauthorized practice of law (UPL) and fiduciary duties particularly enlightening [13, 16]. This body of work stems from similar concerns around individuals relying on the unprotected advice of non-lawyers, risking detrimental real-world outcomes. The insights from this legal perspective shed light on why even when analyzing general-purpose AI systems, domain expertise remains crucial.

Future Work: More Domains, More Democratic Process. We aim to expand this process across domains (mental health, medical advice), distill more dimensions from domain experts, and utilize a democratic process to reach consensus. In a democracy, even when groups disagree, the legitimacy of the process itself means that groups accept the final decision knowing that their opinions were represented. Since policy is defined by judgments on cases, it is important for the community to agree on case decisions. Thus, our final phase focuses on ensuring that precedent decisions involve affected stakeholders, whether the general public or a specific community. As initial steps, we are crowdsourcing public judgements on AI responses to cases.

This research shows how time-tested tools in legal jurisprudence—case-based reasoning—could be adapted to make community norms more tangible and actionable for AI systems. Just as accumulating precedents incrementally advance the law, judgments on use cases can incrementally steer AI towards greater alignment with people's values. By transparently embedding societal perspectives within case repositories, we enable organic governance encouraging responsible innovation. Much work lies ahead in translating identified factors into implementable mechanisms and scaling deliberative case evaluation.

3 Rethinking Concepts of Individual Liberty

Increasingly Textualized World and Free Speech. The rise of algorithmically-mediated communication poses fundamental challenges to traditional distinctions in free speech law between ideas, expression, and conduct. As human activities and experiences occur increasingly through code and programming languages, the scope of *expression* expands as in Figure 4. My scholarship examines how law and rights can evolve amidst this increasingly *textualized* digital world.

Figure 4: Expansion of Expression.

Specifically, my work [6] reveals that the formalistic approach to the First Amendment developed by the U.S. Supreme Court since the 1970s struggles to address the complexities of algorithmic speech. Treating all speech as equally sacrosanct, regardless of speaker identity or industry

structure, incentivizes the weaponization of free speech claims by powerful technology companies. However, I argue that the Court previously employed a more *contextualized* analysis in the early twentieth century, considering factors like distributional outcomes, access inequities, and diversity of voices. This nuanced approach prevents over-expansion of the First Amendment while still safeguarding expressions closely linked to human flourishing.

Privacy Notions after Cambridge Analytica Data Breach. The erosion of privacy in the digital realm has become a pressing issue, threatening historic norms and democratic values. Unlike the physical world where privacy was passively protected by the friction of snooping, the online environment allows for effortless data gathering, making personal information vulnerable to misuse [15]. The Cambridge Analytica data breach in 2016 serves as a stark example of how easily users' privacy can be exploited for non-consensual purposes, in this case, political advertising, ultimately undermining democratic processes.

My research, published in a trilogy of articles within one of the most heavily-cited Korean law reviews, meticulously analyzes the multifaceted response of the three branches of government in the United States to this unprecedented incident. I delve into: (1) **Legislative Initiatives**: examining the California Consumer Privacy Act (CCPA), affording consumers new rights to access, delete and stop the sale of their personal data [3]; (2) **Executive Actions**: analyzing the Federal Trade Commission's (FTC) unprecedented $5 billion settlement with Facebook, highlighting the agency's proactive interpretation of existing laws to curb data misuse [5]; and (3) **Judicial Interpretations**: exploring how federal courts have applied a generous interpretation of "injury in fact" to certify class action lawsuits and expand standing requirements, thereby empowering individuals to seek legal redress for privacy violations [4].

Future Work: Envisioning Cognitive Liberty. While researching Constitutional cornerstones like freedom of expression and privacy, their limitations become apparent when confronted with the subtle manipulations inherent in AI-driven technologies. Existing frameworks focus on protecting individuals from *government interference* and preventing *tangible harm*. However, this approach fails to address the manipulative potential of AI systems operating within the *private sector*. These systems can nudge beliefs, shape choices, and inflict *intangible harms* that are difficult to quantify or link to specific actions.

Recent developments offer promising glimpses into the future. The proposed EU Digital Services Act, with its opt-out provision for algorithmic recommendations, empowers users to exert more control over their online experiences. This aligns with the emerging legal concept of *cognitive liberty*, which conceptualizes the right of individuals to exercise agency and control over their own mental lives as crucial for human flourishing [11]. However, significant work remains to fully realize the promise of cognitive liberty, which includes (1) **addressing how algorithmic interventions influence users' beliefs and conduct**, drawing upon research such as [14]; and (2) **designing legal safeguards** such as transparency requirements, interactive opt-out mechanisms, and algorithmic auditing. Through concrete legal analysis combined with qualitative assessments of human-AI interactions, I aim to advance the understanding of cognitive liberty and develop effective strategies for protecting it.

References

[1] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[2] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. Managing AI Risks in an Era of Rapid Progress. *arXiv preprint arXiv:2310.17688*, 2023.

[3] Inyoung Cheong. After Facebook's 2016 Data Breach I: The California Consumer Privacy Act of 2018. *The Journal of Law and Economic Regulation*, 12(2), 2020.

[4] Inyoung Cheong. After Facebook's 2016 Data Breach II: The U.S. Federal Courts' Consumer Class Actions. *The Journal of Law and Economic Regulation*, 13(1), 2020.

[5] Inyoung Cheong. After Facebook's 2016 data breach III: The FTC's 5-billion dollar settlement. *The Journal of Law and Economic Regulation*, 13(2), 2021.

[6] Inyoung Cheong. Freedom of Algorithmic Expression. *U. Cin. L. Rev.*, 91:680, 2022.

[7] Inyoung Cheong, Aylin Caliskan, and Tadayoshi Kohno. Envisioning Legal Mitigations for Intentional and Unintentional Harms Associated with LLMs. `https://genlaw.github.io/CameraReady/32.pdf`, 2023. presented at ICML 2023 Workshop on Generative AI and Law.

[8] Inyoung Cheong, Aylin Caliskan, and Tadayoshi Kohno. Safeguarding Human Values: Rethinking US Law for Generative AI's Societal Impacts. `https://inyoungcheong.github.io/assets/pdf/AIandEthics.pdf`, 2024. in revision, AI and Ethics.

[9] Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. `https://arxiv.org/abs/2402.01864`, 2024. *arXiv preprint arXiv: 2402.01864*.

[10] Tyna Eloundou and Teddy Lee. Democratic inputs to ai grant program: lessons learned and implementation plans. `https://openai.com/blog/democratic-inputs-to-ai-grant-program-update`, January 2024.

[11] Nita A Farahany. *The battle for your brain: defending the right to think freely in the age of neurotechnology*. St. Martin's Press, 2023.

[12] K. J. Kevin Feng, Quan Ze Chen, Inyoung Cheong, King Xia, and Amy X. Zhang. Case repositories: Towards case-based reasoning for ai alignment. `https://arxiv.org/abs/2310.07019`, 2023. *arXiv preprint: 2311.10934*, presented at NeurIPS 2023 MP2 Workshop.

[13] Claudia E. Haupt. Artificial professional advice. *Yale JL & Tech.*, 21:55, 2019.

[14] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, page 22, New York, NY, USA, 2023. ACM.

[15] Lawrence Lessig. *Free Culture: The Nature and Future of Creativity*. Penguin Press, 2005.

[16] Katherine Medianik. Artificially intelligent lawyers: updating the model rules of professional conduct in accordance with the new technological era. *Cardozo L. Rev.*, 39:1497, 2017.

[17] Wojciech Zaremba, Arka Dhar, Lama Ahmad, Tyna Eloundou, Shibani Santurkar, Sandhini Agarwal, and Jade Leung. Democratic inputs to AI. `https://openai.com/blog/democratic-inputs-to-ai`, May 2023.