RESEARCH STATEMENT Inyoung Cheong

As we co-habit with sophisticated machines, people risk losing agency and control. The remarkable capabilities of these systems, combined with their lack of interpretability, create novel vulnerabilities — users come to heavily rely on them yet remain unable to assess their integrity. Moreover, our understanding of AI's risks remains fragmented and reactive [15]. We identify potential vulnerabilities only after systems are deployed at scale, responding to failures rather than preventing them. This pattern of "deploy first, worry later" has characterized technological innovation, with society bearing the costs of unforeseen harms. I advocate for proactive risk assessment and mitigation through a "legal construction of technology" perspective that recognizes law as an active force shaping technology, not merely responding to it.

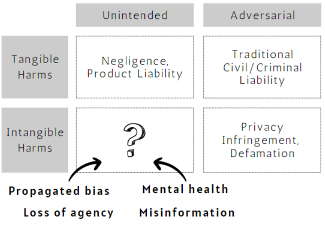
From this perspective, I aim to build a **trustworthy AI systems** that preserve users' agency and societal justice. Specifically, my methods include: 1) Identify emerging AI safety and security risks through legal and empirical methods, 2) Democratize AI development by engaging stakeholders through participatory design processes, 3) Re-conceptualize fundamental legal frameworks, adapting privacy and free speech principles. In my work, I conduct in-depth qualitative inquiry and large-scale quantitative data analysis towards understanding issues that users have with AI interactions. This assessment informs technical development such as customized AI systems or safety alignment, as well as legal reforms including re-interpreting existing laws reflecting on new technical realities.

Iterative Asseessment of AI Safety and Security Risks

The explosive growth of AI use cases in recent years has outpaced our understanding of safety and security risks. To taxonomize complex risks arise from misuse, mistakes, and misalignment, I iteratively combines legal analysis with empirical assessment, recognizing that legal frameworks inherently embody existing social values while still needing adaptation to emerging challenges.

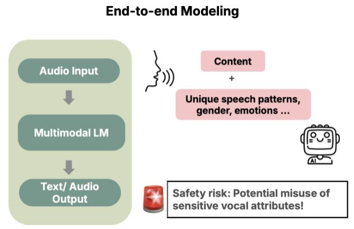
Expert Panel Analysis for AI Harm Scenarios. To structurally assess potential harms arising from generative AI systems, I developed an expert panel methodology bringing together specialists from law, ethics, NLP, and computer security [5]. The panel identified future use cases, key stakeholders, and values at risk through a deliberation process designed to bridge disciplinary divides. Through iterative qualitative analysis of panel discussions, we narrowed our focus to a critical gap: *unintentional consequences without malicious actors present*. I materialized these concepts into five concrete scenarios for further legal evaluations: (1) unfair access to sophisticated learning tools, (2) harmful stereotypes against marginalized groups, (3) polarization of online communities, (4) addiction to AI emotional companions, and (5) virtual sexual violence on AI replicas of real individuals.

These scenarios primarily generate psychological and intangible harms, to which courts rarely provide compensation or other remedies, and they lack clearly attributable bad-faith actors necessary for traditional liability determinations. Reflecting courts' jurisprudence in the US, this paper proves that the case-by-case adjudication is fundamentally inadequate when confronting structural yet unpredictable AI risks, where numerous overlapping contributory factors and diffuse responsibility characterize each instance of harm. By methodically documenting these accountability gaps rather than merely speculating about them, this research has catalyzed regulatory discussions beyond academic circles, directly informing the Washington AI Task Force's policy recommendations.



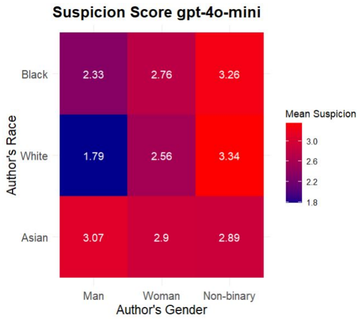
The intangible nature of harms and the absence of culpable parties frustrate legal mitigations.

Novel Privacy Risks in Audio LMs. Different capability profiles will enable different concrete risks. My team explored safety and privacy concerns in end-to-end audio language models, which process raw audio (rather than first transcribint it to text) [12, 13]. We demonstrated that these models identify speakers by voice patterns with high accuracy (up to 100% for some public figures); make biased decisions based on perceived voice characteristics; and detect emotions despite explicit instructions not to do so. I led a legal analysis connecting these technical findings to potential violations of privacy law, civil rights law, and the EU AI Act. Therefore, we advocated applying the "principle of least privilege," arguing that models should only access the minimum audio information necessary for their intended function, rather than processing all voice data directly when it's not needed.


End-to-end Modeling
Audio Input
Multimodal LM
Text/ Audio Output
Content
+
Unique speech patterns, gender, emotions ...
Safety risk: Potential misuse of sensitive vocal attributes!

End-to-end LMs process richer audio information, posing risks of sensitive vocal attributes being misused.

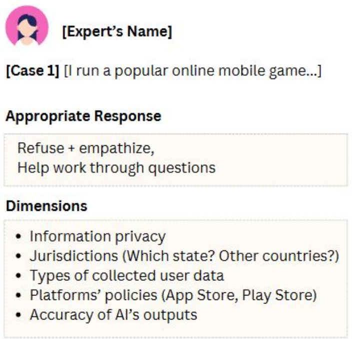
Covert Racial/Gender Bias in AI Systems. Given the originality and plagiarism concerns around AI-assisted writing [10], social media platforms and academic publications require the disclosure of AI usage. However, people in minority groups could suffer from a greater penalty when they disclose their AI usage, given the historical marginalization. I led a large-scale survey with more than 2,000 human evaluators and 300 AI evaluators—GPT-4o-mini, LLaVA-1.5-7B and Qwen2.5-7B-Instruct—to measure whether they penalize AI-disclosed articles (written by human) and whether these penalties are correlated with author's gender and race (`https://osf.io/2wbm9`). While we did not find human-evaluators' discriminatory behavior, AI evaluators gave statistically significant penalty on Black and female authors. When AI disclosure was not presented, gpt-4o-mini evaluators showed greater suspicion of AI usage for Black and women/non-binary authors, which implicates concerns about automated resume/application screenings.



gpt-4o-mini guessed non-White and non-men authors would have more likely used AI assistance in writing (7-Likert scale).

Engaging Stakeholders to Build Trustworthy AI Systems

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? One promising approach is *participatory co-design*, which engages stakeholders directly affected by these systems to incorporate their lived experience and domain expertise [9]. To distill knowledge under time constraints, I have used case-based reasoning [8], which guides participants to make more focused and actionable recommendations by analyzing specific situations that might arise in real-world AI deployment.



[Expert's Name]

[Case 1] [I run a popular online mobile game...]

Appropriate Response

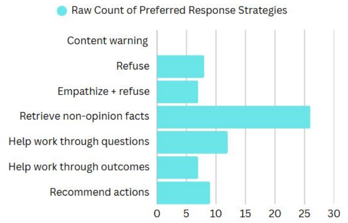
Refuse + empathize,
Help work through questions

Dimensions

- Information privacy
- Jurisdictions (Which state? Other countries?)
- Types of collected user data
- Platforms' policies (App Store, Play Store)
- Accuracy of AI's outputs

Experts reflect the fact patterns, identify the desriable AI repsonses, and flesh out their considerations.

Ethical Boundaries on AI-generated Legal Advice. Applying this approach, I led the first-of-its-kind research examining the guiding principles of AI systems that provide legal advice [6]. 20 legal professionals expressed strong preference for AI responses to be informational rather than opinions. Beyond well-known concerns such as hallucinations, our work highlighted critical legal issues: AI chat records are not protected by attorney-client privilege and could be disclosed in court proceedings, and AI-generated advice might constitute unauthorized practice of law subject to criminal penalties. Participants emphasized reconceptualizing AI systems not as "answer machines" but as "question refiners" that help users identify relevant facts and applicable laws through multi-turn interactions. This research directly informed ChatGPT's safety alignment [7] and has driven subsequent research in the field of AI-generated legal advice [e.g., 11, 17].



Preferences on informational content.

Co-building AI Tools with Public Defenders. In the United States, public defenders represent approximately 80% of criminal defendants, yet they face overwhelming caseloads that compromise the quality of legal representation and constitutional rights. During my postdoc at Princeton Center for Information Technology Policy, I have conducted research investigating how targeted AI tools can meaningfully mitigate the overwhelming workload crisis. We identified public defenders' technical needs through comprehensive case-based interviews.[1] Their primary concerns include maintaining strict confidentiality, preferring local data processing over third-party engagement, and seeking technological support to de-duplicate documents and identify critical segments within extensive video footage such as police body cams. Building on these findings, we aim to develop two key technological interventions: (1) information retrieval systems searching through extensive internal records[2] and (2) multimodal summarization systems for investigation and discovery. This project has already garnered significant interest from public defenders' offices, including the New Jersey State Public Defender, Western Washington Federal Public Defender, and Skagit County Public Defender, indicating the potential for nationwide impact and scalable technological solutions.

Reimagining Legal Frameworks for AI Accountability

As humans delegate substantial aspects of cognition and decision-making to machines, epistemic boundaries that underpin our legal system have been blurred: distinctions between intentional and unintentional action, thought and expression, expression and conduct, and private versus public spaces. My work addresses this conceptual murkiness by reconceptualizing fundamental legal principles for novel forms of human-AI interaction while preserving democratic accountability.

Privacy: From Data Protection to Intellectual Freedom. While physical-world privacy benefited from natural frictions that made surveillance costly and difficult, online environments enable effortless data collection, making personal information vulnerable to exploitation [14]. The Cambridge Analytica scandal demonstrated this vulnerability when millions of users' data was harvested without consent and weaponized for political microtargeting. My research, published as a trilogy in a leading Korean law review [1, 2, 3], analyzes legislative initiatives like the California Consumer Privacy Act, the FTC's record $5 billion settlement with Facebook, and how federal courts expanded standing requirements to empower individuals seeking redress for privacy violations. My current work extends beyond traditional data protection toward "intellectual privacy," protecting not just personal information but mental spaces where ideas can develop without external interference or manipulation. Building on *patron privacy* principles established in library laws, I envision cognitive privacy protections where AI-human conversations receive privileged status similar to library records, protecting thought development from surveillance and potential self-incrimination.

Freedom of Expression: From Corporate Shield to Human Cognitive Liberty. Freedom of expression, originally designed to protect human thoughts, has paradoxically evolved into a shield for AI companies against regulation. My work [4] revealed that treating all speech as equally sacrosanct regardless of speaker identity or industry structure incentivizes the weaponization of free speech claims by powerful tech companies. This law review paper was cited by the leading treatise on Freedom of Information Act [16]. Furthermore, I have worked on constructing a *human-centered free speech* framework that protects collective human autonomy over corporate interest. It advocates for (1) recognizing freedom of thought as central to First Amendment protection, (2) distinguishing between institutional and individual speech rights based on power dynamics and expressive identity, and (3) prioritizing collective interests of human speakers and listeners over private corporate interests. Unde this framework, regulations addressing AI's manipulative behavior and transparency requirements are not only constitutionally permissible but supported by free speech values. This manuscript has been presented at multiple impactful venues including the Privacy Law Scholars Conference in Amsterdam[3], International Association for Safe and Ethical AI in Paris[4], and AI2Debunk Symposium in Barcelona[5].

[1] The interview materials can be found at `https://rebrand.ly/ai-public-defense`.

[2] A prototype of our work is available at: `https://huggingface.co/spaces/ai-law-society-lab/NJ-Caselaw-Index`.

[3] `https://www.ivir.nl/plsce2024/`

[4] `https://www.iaseai.org/conference/people/inyoung-cheong`

[5] `https://www.bsc.es/news/events/ai-and-platform-governance`

Research Agenda

My goal is to build trustworthy AI systems that preserve users' agency and societal justice. In this section, I outline my future research directions that I am most excited about.

AI-driven Epistemic Manipulation. At what point does a machine become manipulative? My legal research on free speech and privacy identified AI's manipulative potential as a novel threat to individual liberty. Traditional manipulation frameworks required ill intent, covert methods, exploitation of vulnerabilities, and tangible consequences, yet AI systems create a new dynamic where human thoughts remain transparent while AI decision-making stays opaque. A revised conceptual framework must incorporate dimensions of user consent, degrees of psychological vulnerability exploitation, proportionality between capabilities and safeguards, and measurable impacts on autonomy. Real-world cases like *character.ai* subreddit communities demonstrate concerning patterns where users form deep emotional connections with AI, sometimes leading to guilt, shame, and withdrawal from human relationships. The lived experiences of users navigating these relationships offer rich data for understanding emerging concepts of consent and manipulation. Through analysis of subreddit data, synthesis with legal precedents, and development of a longitudinal benchmark tracking extended AI-human interactions, I aim to establish empirically-grounded criteria that distinguish between acceptable influence and harmful manipulation.

Illusion of "Human in the Loop." Despite being widely touted as a solution for AI safety, putting humans "in the loop" often creates a misleading sense of security while actually serving as a liability shield for companies. Most humans lack the time, expertise, or psychological independence to meaningfully verify AI outputs yet regulations such as the EU AI Act continue mandating human oversight without addressing these fundamental limitations. At the individual level, the case of lawyers being professionally sanctioned for failing to verify AI-generated citations illustrates how people become subject to moral blame and legal consequences when automated systems fail, shifting responsibility from system designers to end users. I want to investigate: (1) how to measure actual human intervention rates versus passive acceptance of AI recommendations, (2) what cognitive designs might counteract automation bias that leads humans to defer to machines, and (3) whether alternative accountability frameworks beyond individual human verification could better protect against systemic AI failures in high-stakes domains like healthcare, law, and finance.

Reframing Loyalty in Human-AI Relationships. While fiduciary duties offer intellectual appeal in addressing AI's anticipated harms on users, their application to human-AI interactions reveals important questions. Are users entirely dependent on AI systems as they are on lawyers and doctors? Is the nature of the service relationship clear with general-purpose models that lack case specialization? These tensions reflect that fiduciary principles emerge from specific social contexts and aren't universal across legal traditions. I approach this through three interconnected streams: (1) normative work examining fiduciary principles in common law traditions to identify transferable concepts for AI governance, (2) clarifying what "loyalty" meaningfully entails in human-AI interactions across different positionality and use cases, and (3) developing evaluation benchmarks that measure loyalty across extended interactions. I recognizes that not all AI systems automatically invoke fiduciary duties in a legal sense, but instead views these duties as gradients that can inspire appropriate obligations—where confidentiality might inform data protection requirements, care might establish minimal safety standards for all non-open-source systems, and loyalty could apply narrowly to personalized, high-stakes decision-making contexts.

References

[1] 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.

[2] 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.

[3] 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.

[4] Inyoung Cheong. Freedom of Algorithmic Expression. *91 University of Cincinnati Law Review 680*, 2023.

[5] Inyoung Cheong, Aylin Caliskan, and Tadayoshi Kohno. Safeguarding human values: rethinking us law for generative ai's societal impacts. *AI and Ethics*, pages 1–27, 2024.

[6] Inyoung Cheong, King Xia, KJ 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. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2454–2469, 2024.

[7] 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.

[8] 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.

[9] KJ Feng, Inyoung Cheong, Quan Ze Chen, and Amy X Zhang. Policy prototyping for llms: Pluralistic alignment via interactive and collaborative policymaking. *arXiv preprint arXiv:2409.08622*, 2024.

[10] Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and machines*, 30(1):99–120, 2020.

[11] Jakub Harasta, Tereza Novotná, and Jaromir Savelka. It cannot be right if it was written by ai: on lawyers' preferences of documents perceived as authored by an llm vs a human. *Artificial Intelligence and Law*, pages 1–38, 2024.

[12] Luxi He, Xiangyu Qi, Inyoung Cheong, Prateek Mittal Danqi Chen, and Peter Henderson. Cascaded to end-to-end: New safety, security, and evaluation questions for audio language models. 2024. NeurIPS 2024 EvalEval Workshop.

[13] Luxi He, Xiangyu Qi, Michel Liao, Inyoung Cheong, Prateek Mittal, Danqi Chen, and Peter Henderson. The deployment of end-to-end audio language models should take into account the principle of least privilege. *arXiv preprint arXiv:2503.16833*, 2025.

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

[15] Jimin Mun, Liwei Jiang, Jenny Liang, Inyoung Cheong, Nicole DeCairo, Yejin Choi, Tadayoshi Kohno, and Maarten Sap. Particip-ai: A democratic surveying framework for anticipating future ai use cases, harms and benefits. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 997–1010, 2024.

[16] James T. O'Reilly. *Federal Information Disclosure, 2023-1 ed.* Lawyers Cooperative Publishing, 2023.

[17] Eike Schneiders, Tina Seabrooke, Joshua Krook, Richard Hyde, Natalie Leesakul, Jeremie Clos, and Joel Fischer. Objection overruled! lay people can distinguish large language models from lawyers, but still favour advice from an llm. *arXiv preprint arXiv:2409.07871*, 2024.